(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 November 2021 (25.11.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/237221 A1

(51) International Patent Classification:
G06F 9/44 (2018.01) *H04L 12/24* (2006.01)

(21) International Application Number:
PCT/US2021/033942

(22) International Filing Date:
24 May 2021 (24.05.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/029,264 22 May 2020 (22.05.2020) US

(72) Inventors; and
(71) Applicants: **RAO, Shishir R.** [US/US]; 689 Fenley Avenue, San Jose, CA 95117 (US). **RAO, Ravindra Jn** [US/US]; 689 Fenley Avenue, San Jose, CA 95117 (US).

(74) Agent: **BORDERS, Nina Habib**; Reed Smith LLP, 1841 Page Mill Road, Suite 110, Palo Alto, California 94304 (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, IT, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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).

Published:
— *with international search report (Art. 21(3))*

(54) **Title:** MACHINE LEARNING BASED APPLICATION SIZING ENGINE FOR INTELLIGENT INFRASTRUCTURE ORCHESTRATION

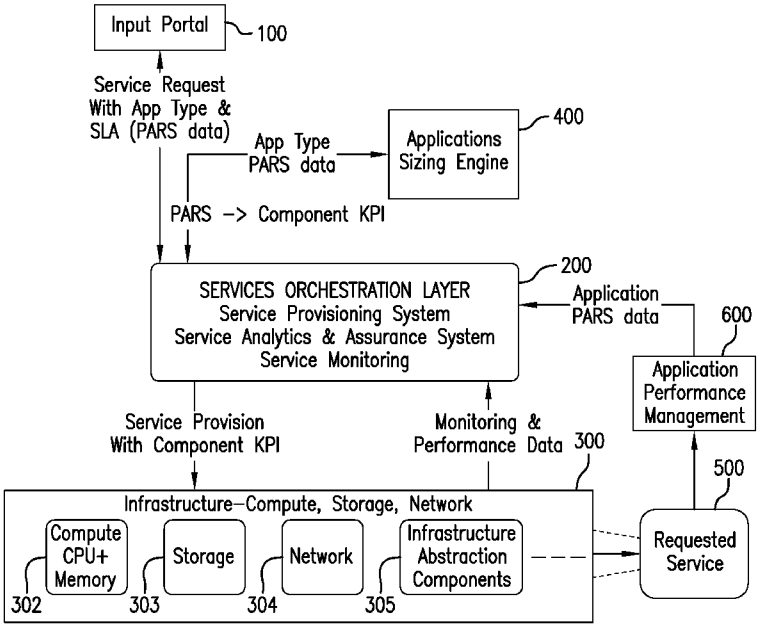

FIG. 1

(57) **Abstract:** This disclosure provides an apparatus, a method and a nontransitory storage medium having computer readable instructions for sizing infrastructure needed for an application as a service.

WO 2021/237221 A1

TITLE

Machine Learning Based Application Sizing Engine
For Intelligent Infrastructure Orchestration

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of U.S. Provisional Patent Application Serial No. 63/029,264 filed on May 22, 2020 under 35 U.S.C. § 119, the entire disclosure of which is incorporated herein by reference.

TECHNICAL FIELD

[0002] This invention relates to business application infrastructure and, more specifically, to facilitate the service provisioning and delivery among cloud service and data center service customers an appropriately sized capacity of the infrastructure components with each component associated with its Key Performance Indicators (KPIs) based on the Intent of the end user.

BACKGROUND

[0003] Cloud computing refers to the use of dynamically scalable computing resources for providing Information Technology (IT) infrastructure for business applications. The computing resources, often referred to as a "cloud," provide one or more services to users. These services may be categorized according to service types, which may include for examples, applications/software, platforms, infrastructure, virtualization, and servers and data storage. The names of service types are often prepended to the phrase "as-a-Service" such that the delivery of applications/software and infrastructure, as examples, may be referred to as Software-as-a- Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure as a Service (IaaS).

[0004] The term "Infrastructure as a Service" or more simply "IaaS" refers not only to infrastructure services provided by an Infrastructure as a Service provider, but also to a form of service provisioning in which cloud customers contract with IaaS service providers for the online delivery of services provided by the cloud. Cloud service providers manage a public, private, or hybrid cloud infrastructure to facilitate the online delivery of cloud services to one or more cloud customers.

SUMMARY

[0005] This disclosure provides a method of sizing the infrastructure for an application as a service, comprising:

receiving the performance, availability, reliability and security information associated with a request for service;

determining an amount of infrastructure and its corresponding Key Performance Indicators (KPIs) to provide for the service based on an empirical model; and

outputting the amount of infrastructure to a service orchestration system.

[0006] Embodiments include:

[0007] The method further comprising

receiving first information associated with the key performance indicators (KPI) of the service's infrastructure components;

predicting the performance of the infrastructure based on the KPI; receiving second information associated with observed performance of the infrastructure;

comparing the predicted performance based on the KPI with the observed performance;

converting the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

updating the weights of the KPI and performance characteristics using the machine learning algorithm.

[0008] The method further comprising

determining a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics; and

outputting the sizing solution along with the updated KPIs to the service orchestration system.

[0009] This disclosure also provides an apparatus for sizing infrastructure for an application as a service, comprising:

a memory; and

at least one processor coupled to the memory, the processor configured to: receive information associated with a request for service;

determine an amount of infrastructure to provide the service based on an empirical model;

output the amount of infrastructure to a service orchestration systems.

[0010] Embodiments include:

[0011] The apparatus wherein the processor is further configured to

receive first information associated with the key performance indicators (KPI) of the infrastructure components;

predict the performance of the infrastructure based on the KPI;

receive second information associated with observed performance of the infrastructure;

compare the predicted performance based on the KPI with the observed performance;

convert the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

update the weights of the KPI and performance characteristics using the machine learning algorithm.

[0012] The apparatus wherein the processor is further configured to

determine a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics;

output the sizing solution to the service orchestration system.

[0013] This disclosure also provides a non-transitory computer readable medium having computer readable instructions stored thereon, that when executed by a computer cause at least one processor to,

receive information associated with a request for service;

determine an amount of infrastructure to provide the service based on an empirical model; and

output the amount of infrastructure to a service orchestration system.

[0014] Embodiments include:

[0015] The non-transitory computer readable medium wherein the computer readable instructions further cause at least one processor to

receive first information associated with the key performance indicators (KPI) of the infrastructure components;

predict the performance of the infrastructure based on the KPI;

receive second information associated with observed performance of the infrastructure;

compare the predicted performance based on the KPI with the observed performance;

convert the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

update the weights of the KPI and performance characteristics using the machine learning algorithm.

[0016] The non-transitory computer readable medium wherein the computer readable instructions further cause at least one processor to

determine a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics; and

[0017] output the sizing solution to a service orchestration system.

[0018] The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Fig. 1 shows the architecture of the Application Service Provisioning System according to exemplary embodiments of the disclosure.

[0020] Figures 2a and 2B show the user input portal according to exemplary embodiments of the disclosure.

[0021] Figure 3 shows the Service Orchestration System according to exemplary embodiments of the disclosure.

[0022] Figure 4 shows aspects of the infrastructure in the topology of the infrastructure as a service according to exemplary embodiments of the disclosed subject matter.

[0023] Figure 5 shows aspects of the Application Sizing Engine according to exemplary embodiments of the disclosed subject matter.

[0024] Fig 6A shows aspects of the ML based Resource Optimizer operating in a Learning Mode according to exemplary embodiments of the disclosed subject matter.

[0025] Fig 6B shows aspects of the ML based Resource Optimizer operating in a Predict Mode according to exemplary embodiments of the disclosed subject matter.

[0026] Fig 6C shows aspects of the Capacity and KPI Remediation according to exemplary embodiments of the disclosed subject matter.

[0027] Fig 7 shows aspects of the input portal form for a given Application type according to exemplary embodiments of the disclosed subject matter.

[0028] Fig 7a shows aspects of the workflow from the input portal to the infrastructure according to exemplary embodiments of the disclosed subject matter.

[0029] Fig 8 depicts the workflow of how a particular Application Type will be delivered for the very first time on the infrastructure according to exemplary embodiments of the disclosed subject matter.

[0030] Fig 9 depicts the workflow of the ASE in learning mode according to exemplary embodiments of the disclosed subject matter.

[0031] Fig 10 depicts the workflow of the ASE's ML algorithms in the predict mode, according to exemplary embodiments of the disclosed subject matter.

[0032] Fig 11 depicts the workflow of the ASE according to exemplary embodiments of the disclosed subject matter.

BRIEF DEFINITIONS

[0033] P.A.R.S. characteristics: Performance, Availability, Reliability, and Security characteristics include the following parameters.

[0034] Performance characteristics: parameters upon which the application performance is measured. Specifically: Transactions per Second (TPS), number of concurrent transactions, latency per transaction, etc.

[0035] Availability characteristics: measurement of time that defines the availability of the application for a user. Specifically: degree of availability may be defined by the number of "9's" in percentage and Recover Point Objective (RPO). Number of "9's" e.g. "3 9's" means 99.9%

availability of said application, "4 9's" means 99.99% availability of said application and so on. Also, there is the measurement of RPO, measured in seconds, which means: in the event that the service is lost, this a measurement in seconds of the maximal allowance of time lag for which the application will allow.

[0036] Reliability Characteristics: a measure of reliability which is binary and involves allocating/not allocating "(n+1)" resources for the application infrastructure.

[0037] Security Characteristics: the parameters involved for delivering the required level of security for said infrastructure. Specifically, the level of privacy of infrastructure in terms of infrastructure resources and hardware.

[0038] Input Output Operations Per Second (IOPS): The number of input and output operations per second, one of the performance parameters for disk storage systems.

[0039] Key Performance Indicators (KPI): performance characteristics of components (storage, network, memory, and computational components). Specifically these may be measured in percentage utilized of CPU, percentage utilized of memory components, latency and IOPS of storage components, maximum bandwidth required and error rate of network components, etc.

[0040] Service Level Agreement (SLA) – P.A.R.S. characteristics agreed upon by user and service provider.

[0041] Capacity – The required amount of classified resource (Compute, Storage, Network, etc.) needed to deliver the service.

DETAILED DESCRIPTION OF ILLUSTRATED EMBODIMENTS

[0042] The following detailed description of example implementations refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements.

[0043] When a user, henceforth referred to as U requires compute, memory, storage, and network services to host and maintain a certain business application, U will request a service provider to accurately provision the said Infrastructure as a Service. The Application Sizing Engine, henceforth referred to as ASE, aims to calculate the amount of individual components and provision U with components (e.g. compute, memory, storage, and network components) to host said application adhering to SLA between U and Service Provider at all times.

[0044] This document in general describes the Machine Learning (ML) based Application Sizing Engine in detail and other intelligent infrastructure orchestration components in an application service provisioning system. The specific component discussed in depth in this document is the Application Sizing Engine where the said module will facilitate the provisioning of appropriate infrastructure based on the intent provided by the user. This module will make the calculations for successfully provisioning the Infrastructure as a Service for the user's application. This Application Sizing Engine as described herein will, as a result, facilitate provisioning a business-level service according to well-defined service policies, quality of service, service level agreements, and costs, and further according to a service topology for the business-level service.

[0045] The Application Sizing Engine (ASE) comprises a software module that will size the appropriate infrastructure components required to accurately provision the application to satisfy the intent of the application and will communicate with an Infrastructure as a Service Orchestration System to deliver the requisite application infrastructure. The ASE achieves this by first delivering the capacity and Key Performance Indicators (KPIs) of the individual components based on the Application Service Level Agreement (SLA) provided by the user using empirical data. In this time, the ASE will train or teach the Machine Learning (ML) module to learn associations between the infrastructure components, KPI, and finally the Performance characteristics. After validating the ML module's propensity to learn these relationships, the ASE will leverage this trained module to ensure the SLA intended for the application is adhered to by training the data model based out of the current infrastructure and later correcting the predicted capacity and KPI of the component(s) if necessary.

[0046] An Application is a computer program or a group of programs designed to assist in performing a business activity. The application is executed on one of more infrastructure components and the capacity or the number of these components will depend on the complexity of the application. For eg. An online transaction database (OLTP), a data warehousing database (DW), a web server or a messaging server are different application types that can be executed on the infrastructure.

[0047] The SLA differs from application to application and business to business. The SLA is a combination of the PARS parameters defined above. For eg. The SLA of an OLTP database could be;

a. Performance: 2500 Transactions per second, less than 2 Secs latency per transaction and 500 concurrent transactions
b. Availability: 4-9s (Downtime: 52 mins, 36 secs per year)
c. Reliability: Clustered servers for redundancy
d. Security: Independent hardware.

[0048] For a web server the SLA definition will be different as follows:

a. Performance: Load Time less than 1.5 Secs, Speed Index of less than 2500 ms
b. Availability: 5-9s (Downtime of 25 mins 30 secs per year)
c. Reliability None
d. Security: Shared hardware.

[0049] While the infrastructure service provider, henceforth referred to as ISP will provision the service with components that *aim* to meet the requirements of said application, the ASE aims to accurately size the individual components for said application that meet (or exceed) SLA requirements. Initially obtaining the application type and corresponding SLA requirements (PARS parameters) for an application from the Service Orchestration Systems, which is a software module that provides the infrastructure sizing and minimum thresholds for KPIs for the individual infrastructure components to meet SLA requirements. Additionally in runtime, the module will autonomously (using machine learning) resize the infrastructure for said application based on the service analytics and assurances data provided by the Service Analytics and Assurances Systems within the Service Orchestration Systems.

[0050] Figure 1 shows the architecture of the Application Service Provisioning System and consists of following major components:

a. Block 100 : The Input Portal where the user will input the intent for the application type required.
b. Block 200 : The Service Orchestration system that Provisions, monitors, assures and remediates an application service delivered. In addition to these services the block also has other functions such as infrastructure registry and infrastructure services.
c. Block 300 : Described in Figure 4.

d. Block 400 : The Application Sizing Engine which sizes the capacity and performance Key Performance Indicators (KPIs) of the components of the service, further described in Figures 5, 6a, 6b and 6c.

e. Block 500 : The requested service itself, the requested service can be a bare metal server, a Virtual Machine running on a Hypervisor or a Container that hosts the required application.

f. Block 600 : An external Application Performance Management (APM) software which would monitor the requested service to provide the observed performance KPIs. Block 600 can be a commercially available APM software provided by vendors such as Dynatrace, Cisco or New Relic.

[0051] Figures 2a, 2b show the user input portal (block 100) according to exemplary embodiments of the disclosure. Block 100, containing blocks 110, 120, 130 and 140, allows the user to provide intent of the Infrastructure service required for a particular type of application.

[0052] Figure 3 shows the Service Orchestration System (block 200) according to exemplary embodiments of the disclosure.

[0053] Fig 4 shows aspects of the infrastructure in the topology of the infrastructure as a service according to exemplary embodiments of the disclosed subject matter, including Block 300. Infrastructure may contain at least Blocks 310, 320, 330, and/or 340.

a. Block 320: Compute – Physical compute components

b. Block 330: Storage – Physical storage components

c. Block 340: Network – Physical network components

d. Block 350: Infrastructure Abstraction Components

e. Block 360: Operations support functions needed to efficiently run the Infrastructure, viz. DNS, DHCP, NTP, Patch Management, etc.

f. Block 370: Business support functions needed to efficiently run the business, viz. CRM systems, billing systems, etc.

g. Block 380: Operator tools needed to efficiently run the infrastructure, viz. email, pager, messaging channels, help desk, ticketing systems, etc.

h. Block 390: Communications functions needed to communicate with the personnel managing the infrastructure, viz. phone, wireless communication devices, etc.

i. Block 300 will also contain any infrastructure component that is a component of the service needed to be delivered, and this can be extended to the physical attributes like power distribution units, Heating, Ventilation and Air Conditioning (HVAC) systems, etc.

[0054] Figure 5 shows aspects of the Application Sizing Engine according to exemplary embodiments of the disclosed subject matter. The Application Sizing Engine (ASE) includes Block 400 of the topology of Figure 1 containing Blocks 410, 420, 430, 440, 450, 460 and 470 performing the function of appropriately sizing the infrastructure that needs to be provisioned to as per the intent of the user.

[0055] Fig 6A shows aspects of the ML based Resource Optimizer operating in a Learning Mode. The resource optimizer is training the data set in this mode of operation.

[0056] Fig 6B shows aspects of the ML based Resource Optimizer operating in a Predict Mode. The resource optimizer predicts the correct component capacity and KPIs in this mode of operation.

[0057] Fig 6C shows aspects of the Capacity and KPI Remediation according to exemplary embodiments of the disclosed subject matter. This consists of Block 490, 491, 492 and 493, based on the prediction provided by the Performance Characteristic Prediction module, predicts the new or changed capacity and its corresponding KPIs.

[0058] Fig 7 outlines the input portal form for a given application type and outlines the different components that will be invoked to deliver the requested service.

[0059] Fig 7a outlines the workflow from the input portal to the infrastructure to show how the requested service will be delivered.

[0060] Next is a description of the process of service provisioning to user (U). To accompany said description, there exists an example service provision outline below for an embodiment of the disclosed invention.

[0061] U will approach the user portal – Block 100 – and request Infrastructure Service to be provisioned for an Online Transaction Processing Database (OLTP Database) with capacity of 10 terabytes. U requests that the infrastructure service for said application must meet certain P.A.R.S requirements outlined below:

a. Performance Characteristics – Transactions per Second: 3000; Number of concurrent transactions: 500; Transaction latency: ≤ 1 second (transaction should complete within one second)
b. Availability Characteristics – 99.999% availability (5-9's)
c. Reliability Characteristics – High Availability Enabled
d. Security Characteristics – Dedicated resources, shared hardware

[0062] U will input said P.A.R.S. requirements establishing the SLA between U and SP on the user portal in Block 100. The P.A.R.S. characteristics established by U will be shared with the Intent Based Application Infrastructure as a Service Orchestration System – Block 200. Specifically, the said data will be transmitted to the Service Orchestration System. The Service Orchestration System will communicate the P.A.R.S. characteristics with the ASE – Block 400 to devise a possible solution that meets and adheres to the SLA. This solution involves providing the capacity and the KPIs of the individual components using one of the two methods described below:

a. In the event the particular application type is being deployed by Block 200 for the very first time, the ASE utilizes an already stored empirical model to provide the capacity and KPIs of the individual components
b. In the event that the particular application type has already been deployed by Block 200, then the ASE has already trained its data model for the ML algorithm and it will provide the capacity and KPIs for the individual components based on the current state and performance of the given infrastructure, Block 300 and specifically provision the compute, storage, network and Infrastructure abstraction components – Blocks 302, 303, 304, 305, etc. respectively.

[0063] Fig 8 depicts the workflow of how a particular application type will be delivered for the very first time on the infrastructure using empirical modeling done in advance of a Service Level Request. Figure 8 describes the workflow that is followed by the Blocks 100, 200, 400 and 300 when a particular application type is being deployed for the very first time,

[0064] Block 200 receives the sizing and the KPIs of the required infrastructure components. Block 200 finds the appropriate component within the infrastructure and through the communication medium previously determined between the Service Orchestration System – Block 200 and the individual component. Once the component is configured as per the request, Block

200, the service orchestration system performs all the necessary tasks to ensure that the individual components are all configured to perform as a single application service entity.

[0065] The first time an application type is deployed the ASE will enter the learning mode, the inputs for its learning and training the ML algorithm are the KPIs being observed for the requested service components and Application Performance Management software or an operator manually entering the observed SLA of the requested service. The ASE uses these two inputs to compare and teach the ML algorithms of its previous empirical predictions and retrain the data set to a more accurate predictions based on real time inputs from the infrastructure in Block 300.

[0066] Figure 9 describes the workflow that is followed to train the ML algorithms with real time data, which enables the ML algorithms to learn about the infrastructure and its behavior for a particular application type.

[0067] Once the ASE is trained with the data attributes of the current infrastructure, the ASE operates in two modes:

a. The mode in which it receives real time component KPIs and application performance data from Block 200 about the services of the particular application type it has just been trained on, and now operates to remediate the said requested service to operate at optimal capacity levels

b. The mode in which ASE provides a more accurate sizing of capacity and KPIs for a brand new requested services of the same application type

[0068] Fig 10 depicts the workflow of the ASE's ML algorithms in the predict mode, which enables the ML engine to predict the corrections needed to the current sizing and KPI characteristics based on the performance of the current infrastructure for a particular application type.

[0069] Figure 10 along with Figure 6C highlights the workflow that is used by ASE to provide a recalculated component capacity and KPI for an already existing requested service and ensure that the resources are optimally utilized for a given application type.

[0070] Fig 11 depicts the workflow of the ASE which gets its recommendations from the ML algorithms and initiates Block 200 to take corrective action.

[0071] Figure 11 shows the workflow for a new requested service to be deployed on the infrastructure for the given application type that has already been deployed at least once on the infrastructure.

[0072] The foregoing disclosure provides illustration and description, but is not intended to be exhaustive or to limit the implementations to the precise form disclosed. Modifications and variations are possible in light of the above disclosure or may be acquired from practice of the implementations.

[0073] As used herein, the term component is intended to be broadly construed as hardware, firmware, a combination of hardware and software and/or a particular Information Technology function such as compute, network or storage.

[0074] Certain user interfaces have been described herein and/or shown in the figures. A user interface may include a graphical user interface, a non-graphical user interface, a text-based user interface, etc. A user interface may provide information for display. In some implementations, a user may interact with the information, such as by providing input via an input component of a device that provides the user interface for display. In some implementations, a user interface may be configurable by a device and/or a user (e.g., a user may change the size of the user interface, information provided via the user interface, a position of information provided via the user interface, etc.). Additionally, or alternatively, a user interface may be pre-configured to a standard configuration, a specific configuration based on a type of device on which the user interface is displayed, and/or a set of configurations based on capabilities and/or specifications associated with a device on which the user interface is displayed.

[0075] To the extent the aforementioned embodiments collect, store or employ personal information provided by individuals, it should be understood that such information shall be used in accordance with all applicable laws concerning protection of personal information. Additionally, the collection, storage and use of such information may be subject to consent of the individual to such activity, for example, through well known "opt-in" or "opt-out" processes as may be appropriate for the situation and type of information. Storage and use of personal information may be in an appropriately secure manner reflective of the type of information, for example, through various encryption and anonymization techniques for particularly sensitive information.

[0076] It will be apparent that systems and/or methods, described herein, may be implemented in different forms of hardware, firmware, or a combination of hardware and software. The actual specialized control hardware or software code used to implement these systems and/or methods is not limiting of the implementations. Thus, the operation and behavior of the systems

and/or methods were described herein without reference to specific software code—it being understood that software and hardware can be designed to implement the systems and/or methods based on the description herein.

[0077] Even though particular combinations of features are recited in the claims and/or disclosed in the specification, these combinations are not intended to limit the disclosure of possible implementations. In fact, many of these features may be combined in ways not specifically recited in the claims and/or disclosed in the specification. Although each dependent claim listed below may directly depend on only one claim, the disclosure of possible implementations includes each dependent claim in combination with every other claim in the claim set.

[0078] No element, act, or instruction used herein should be construed as critical or essential unless explicitly described as such. Also, as used herein, the articles "a" and "an" are intended to include one or more items, and may be used interchangeably with "one or more." Furthermore, as used herein, the term "set" is intended to include one or more items, and may be used interchangeably with "one or more." Where only one item is intended, the term "one" or similar language is used. Also, as used herein, the terms "has," "have," "having," or the like are intended to be open-ended terms. Further, the phrase "based on" is intended to mean "based, at least in part, on" unless explicitly stated otherwise.

CLAIMS

We claim:

1. A method of sizing infrastructure for an application as a service, comprising: receiving information associated with a request for service;

determining an amount of infrastructure to provide the service based on an empirical model;

determining the corresponding Key Performance Indicators (KPIs) for the infrastructure based on the empirical model; and

outputting the amount of infrastructure to a service orchestration system.

2. The method of claim 1, further comprising:

receiving first information associated with the key performance indicators (KPI) of the infrastructure components;

predicting the performance of the infrastructure based on the KPI;

receiving second information associated with observed performance of the infrastructure;

comparing the predicted performance based on the KPI with the observed performance;

converting the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

updating the weights of the KPI and performance characteristics using the machine learning algorithm.

3. The method of claim 2, further comprising:

determining a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics; and

outputting the sizing solution to the service orchestration system.

4. An apparatus for sizing infrastructure for an application as a service, comprising:

a memory; and

at least one processor coupled to the memory, the processor configured to:

receive information associated with a request for service;

determine an amount of infrastructure to provide the service based on an empirical model;

determine the corresponding Key Performance Indicators (KPIs) for the infrastructure based on the empirical model; and

output the amount of infrastructure to a service orchestration system.

5. The apparatus of claim 4, wherein the processor is further configured to receive first information associated with the key performance indicators (KPI) of the infrastructure components;

predict the performance of the infrastructure based on the KPI;

receive second information associated with observed performance of the infrastructure;

compare the predicted performance based on the KPI with the observed performance;

convert the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

update the weights of the KPI and performance characteristics using the machine learning algorithm.

6. The apparatus of claim 5, wherein the processor is further configured to

determine a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics; and

output the sizing solution to the service orchestration system.

7. A non-transitory computer readable medium having computer readable instructions stored thereon, that when executed by a computer cause at least one processor to:

receive information associated with a request for service;

determine an amount of infrastructure to provide the service based on an empirical model;

determine the corresponding Key Performance Indicators (KPIs) for the infrastructure based on the empirical model; and

output the amount of infrastructure to a service orchestration system.

8. The non-transitory computer readable medium of Claim 7 wherein the computer readable instructions further cause at least one processor to:

receive first information associated with the key performance indicators (KPI) of the infrastructure components;

predict the performance of the infrastructure based on the KPI;

receive second information associated with observed performance of the infrastructure;

compare the predicted performance based on the KPI with the observed performance;

convert the observed performance, availability, reliability and security parameters of the infrastructure into homogenized space vectors for a machine learning algorithm; and

update the weights of the KPI and performance characteristics using the machine learning algorithm.

9. The non-transitory computer readable medium of Claim 8 wherein the computer readable instructions further cause at least one processor to

determine a sizing solution for an amount of infrastructure to provide the service based on the updated weights of the KPI and performance characteristics; and

output the sizing solution to the service orchestration system.

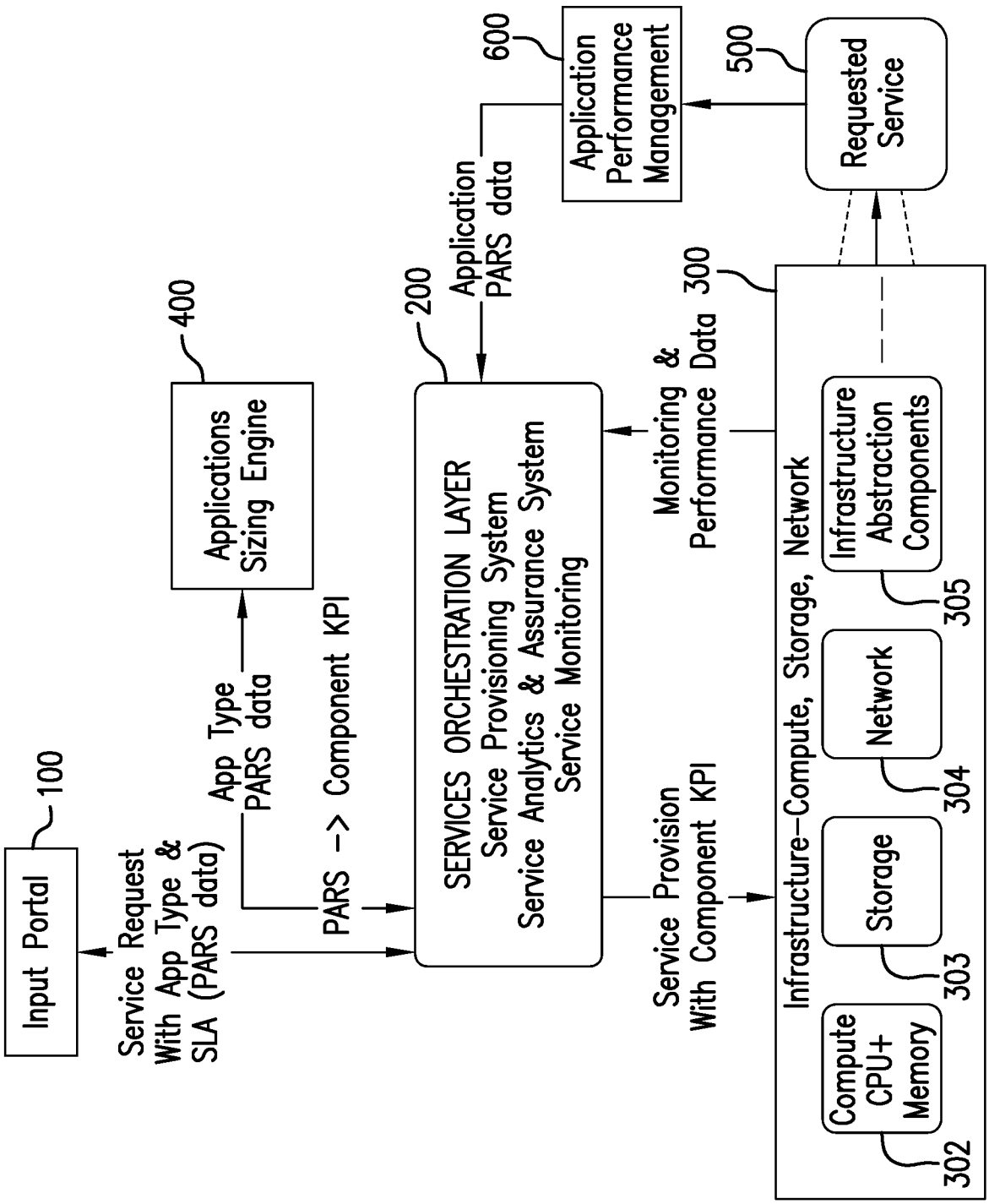
Input Portal
100
Service Request
With App Type &
SLA (PARS data)
App Type
PARS data
Applications
Sizing Engine
400
PARS –> Component KPI
SERVICES ORCHESTRATION LAYER
Service Provisioning System
Service Analytics & Assurance System
Service Monitoring
200
Application
PARS data
Application
Performance
Management
600
Requested
Service
500
Monitoring &
Performance Data
Service Provision
With Component KPI
Infrastructure–Compute, Storage, Network
300
Infrastructure
Abstraction
Components
305
Network
304
Storage
303
Compute
CPU+
Memory
302

FIG. 1

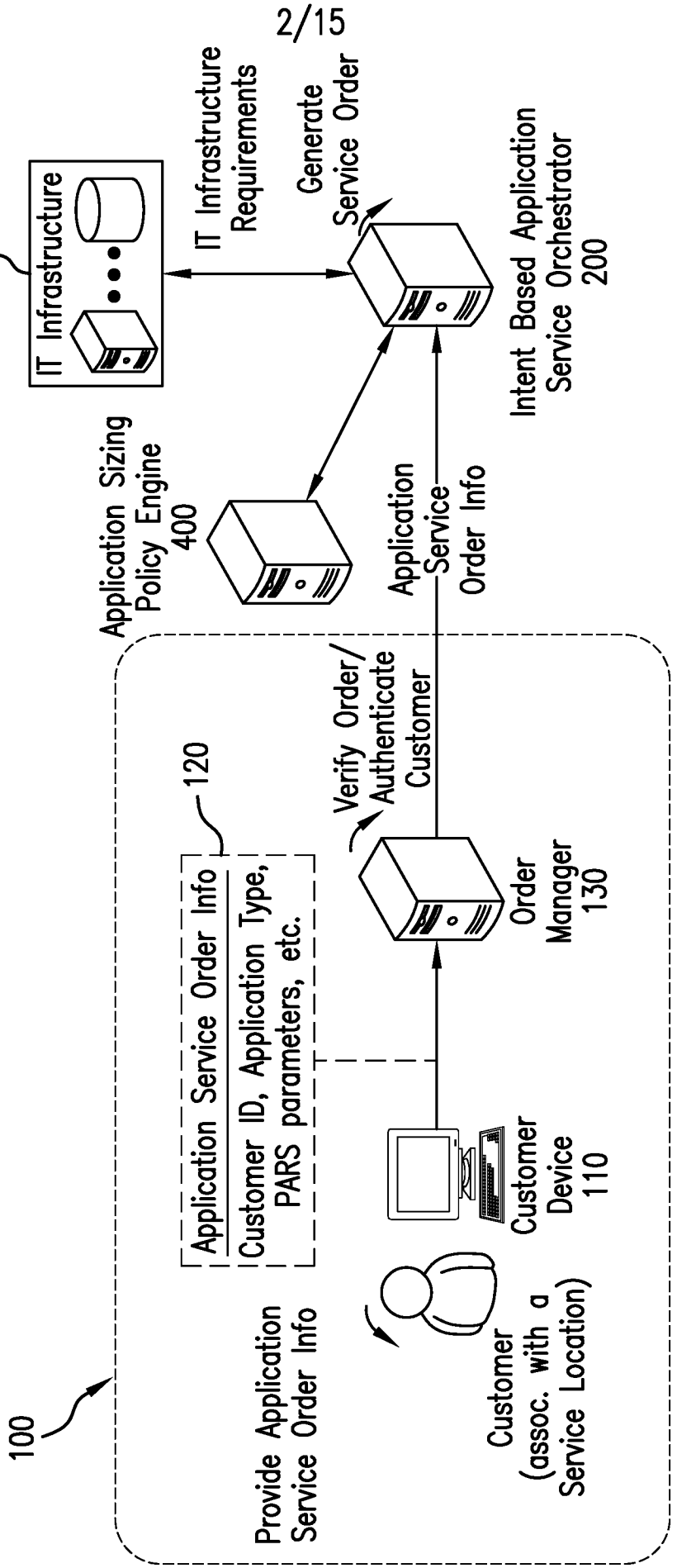
100
Provide Application Service Order Info
Application Service Order Info
Customer ID, Application Type, PARS parameters, etc.
120
Customer (assoc. with a Service Location)
Customer Device 110
Verify Order/ Authenticate Customer
Order Manager 130
Application Service Order Info
Application Sizing Policy Engine 400
IT Infrastructure Requirements
Generate Service Order
Intent Based Application Service Orchestrator 200
300
IT Infrastructure

FIG.2A

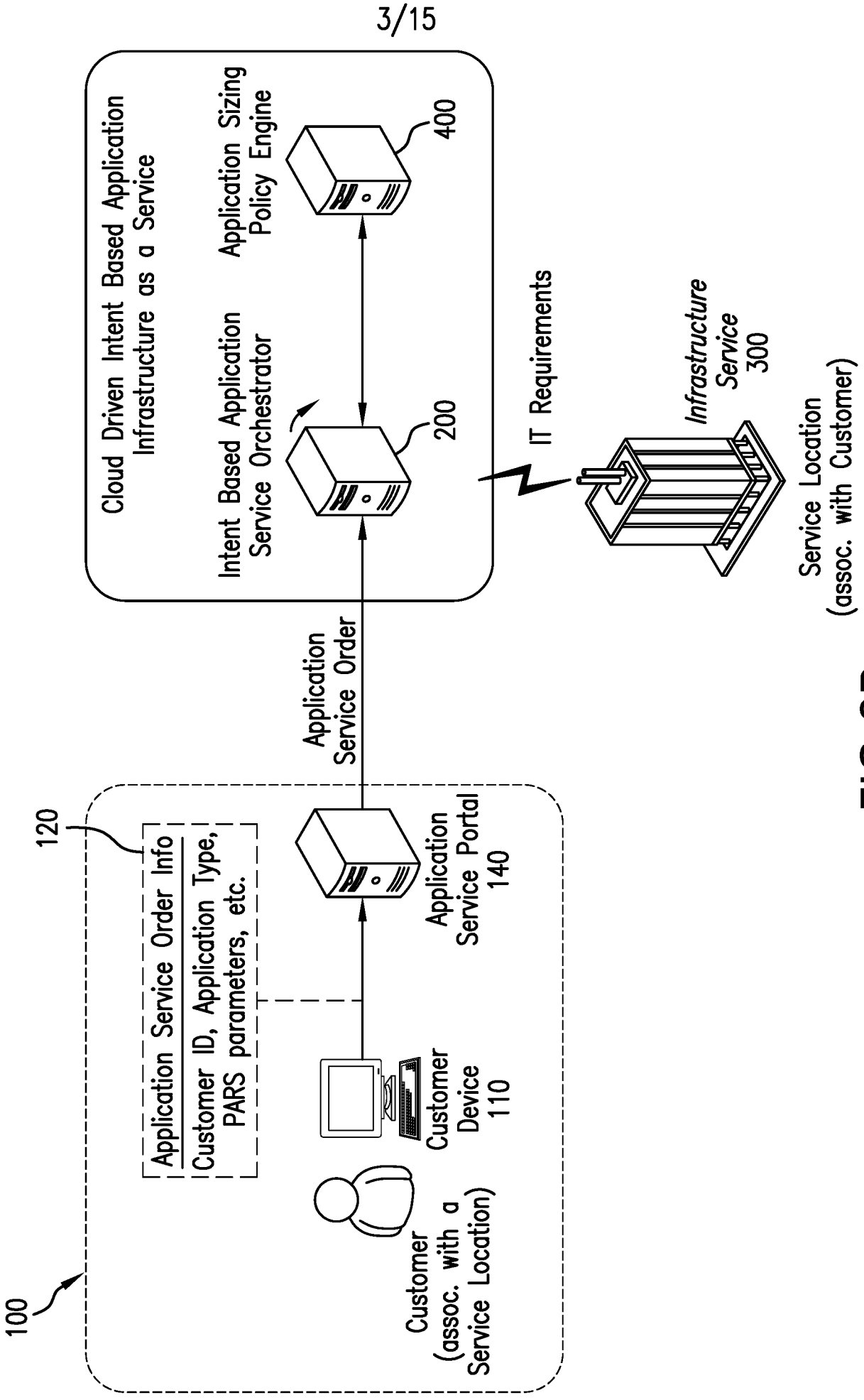
Cloud Driven Intent Based Application Infrastructure as a Service
Application Sizing Policy Engine
400
Intent Based Application Service Orchestrator
200
IT Requirements
Infrastructure Service 300
Service Location (assoc. with Customer)
Application Service Order
120
Application Service Order Info
Customer ID, Application Type, PARS parameters, etc.
Application Service Portal 140
Customer Device 110
Customer (assoc. with a Service Location)
100

FIG.2B

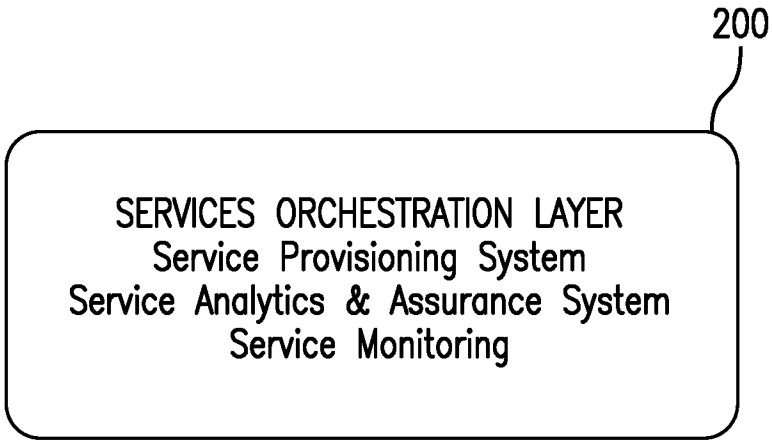
200
SERVICES ORCHESTRATION LAYER
Service Provisioning System
Service Analytics & Assurance System
Service Monitoring

FIG.3

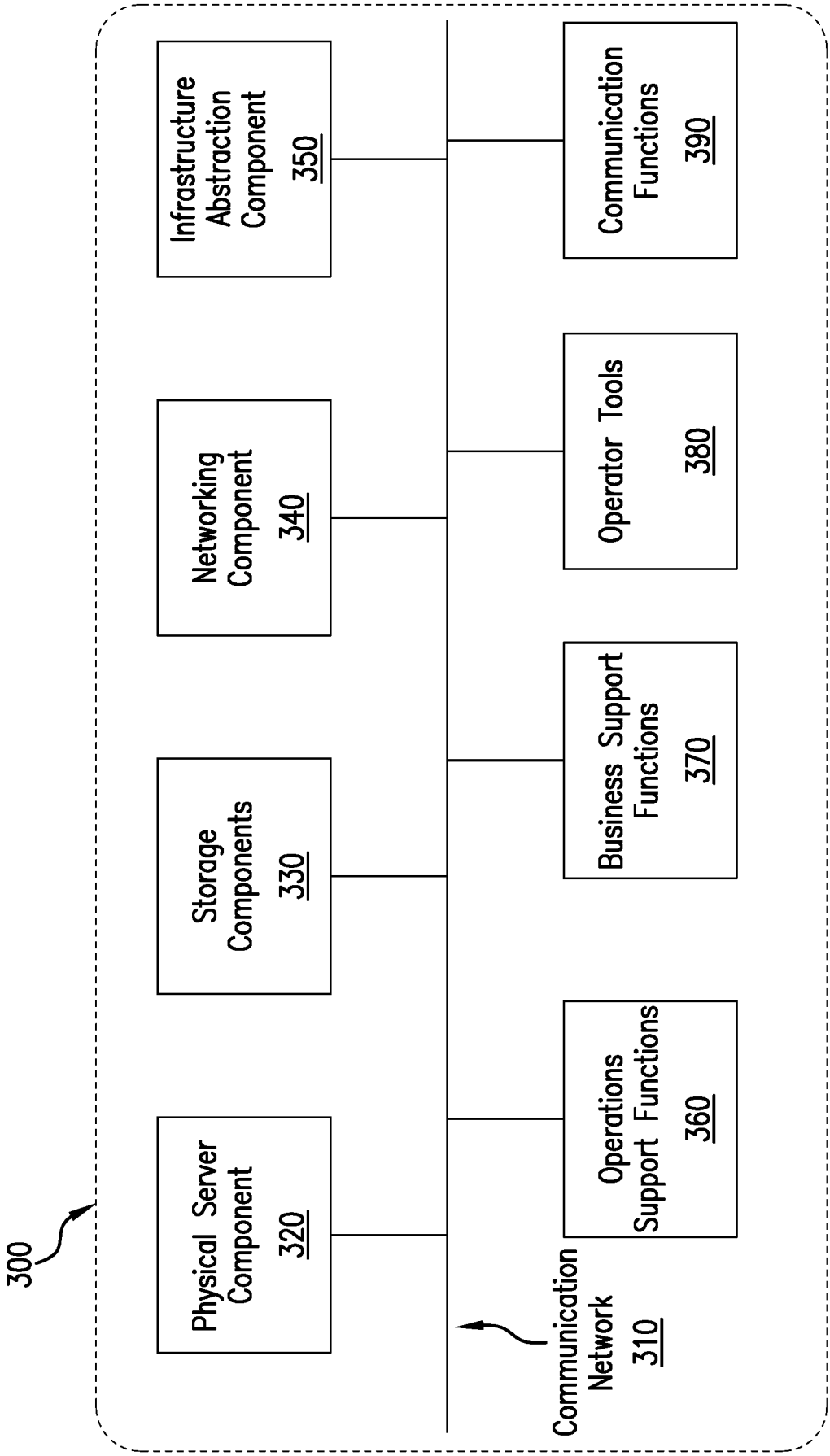
300
Physical Server Component 320
Storage Components 330
Networking Component 340
Infrastructure Abstraction Component 350
Operations Support Functions 360
Business Support Functions 370
Operator Tools 380
Communication Functions 390
Communication Network 310

FIG.4

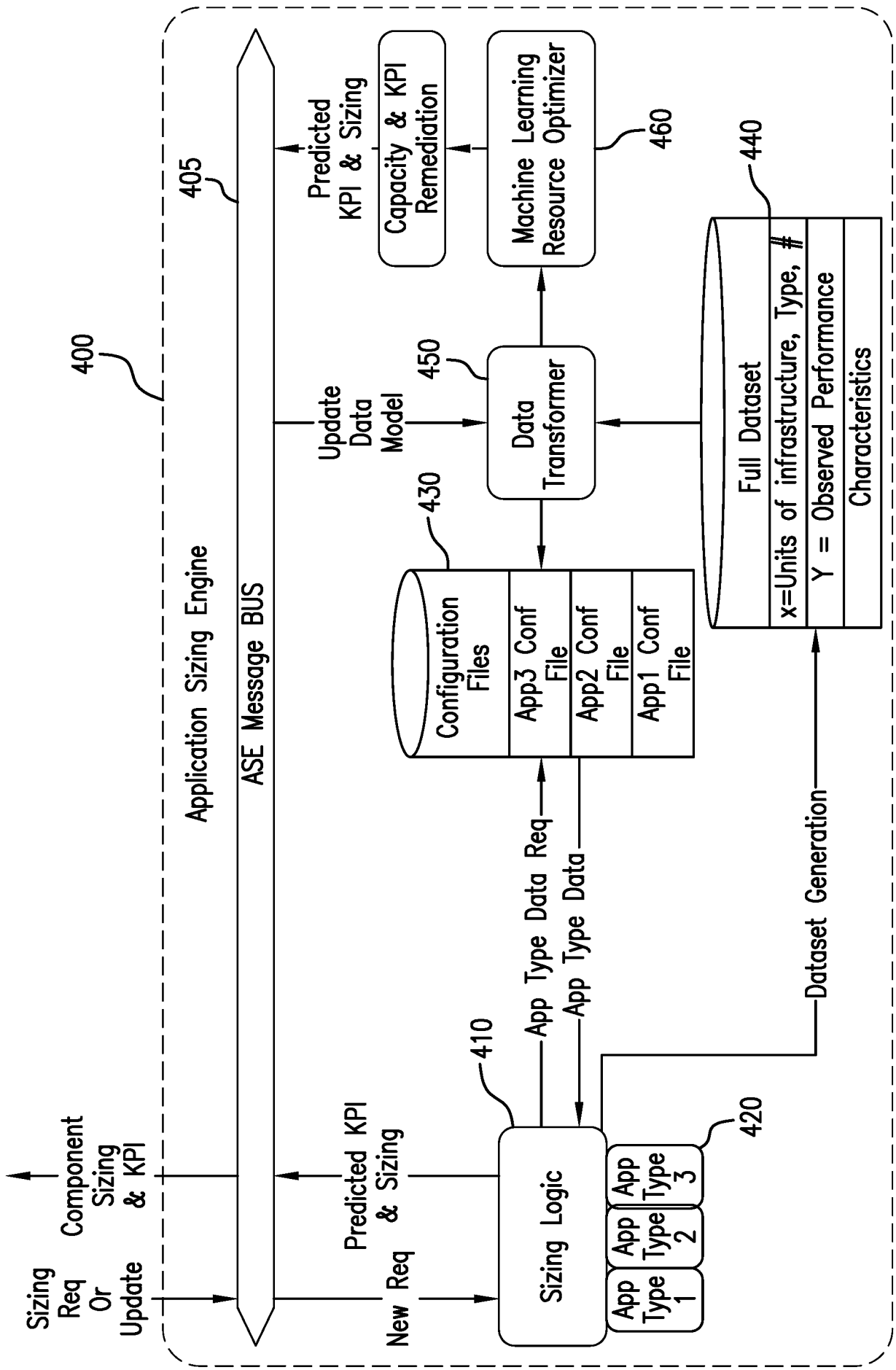
400
Application Sizing Engine
405
ASE Message BUS
Sizing Req Or Update
Component Sizing & KPI
Predicted KPI & Sizing
New Req
410
Sizing Logic
App Type 1
App Type 2
App Type 3
420
App Type Data Req
App Type Data
430
Configuration Files
App3 Conf File
App2 Conf File
App1 Conf File
Update Data Model
450
Data Transformer
Predicted KPI & Sizing
Capacity & KPI Remediation
Machine Learning Resource Optimizer
460
Dataset Generation
440
Full Dataset
x=Units of infrastructure, Type, #
Y = Observed Performance
Characteristics

FIG.5

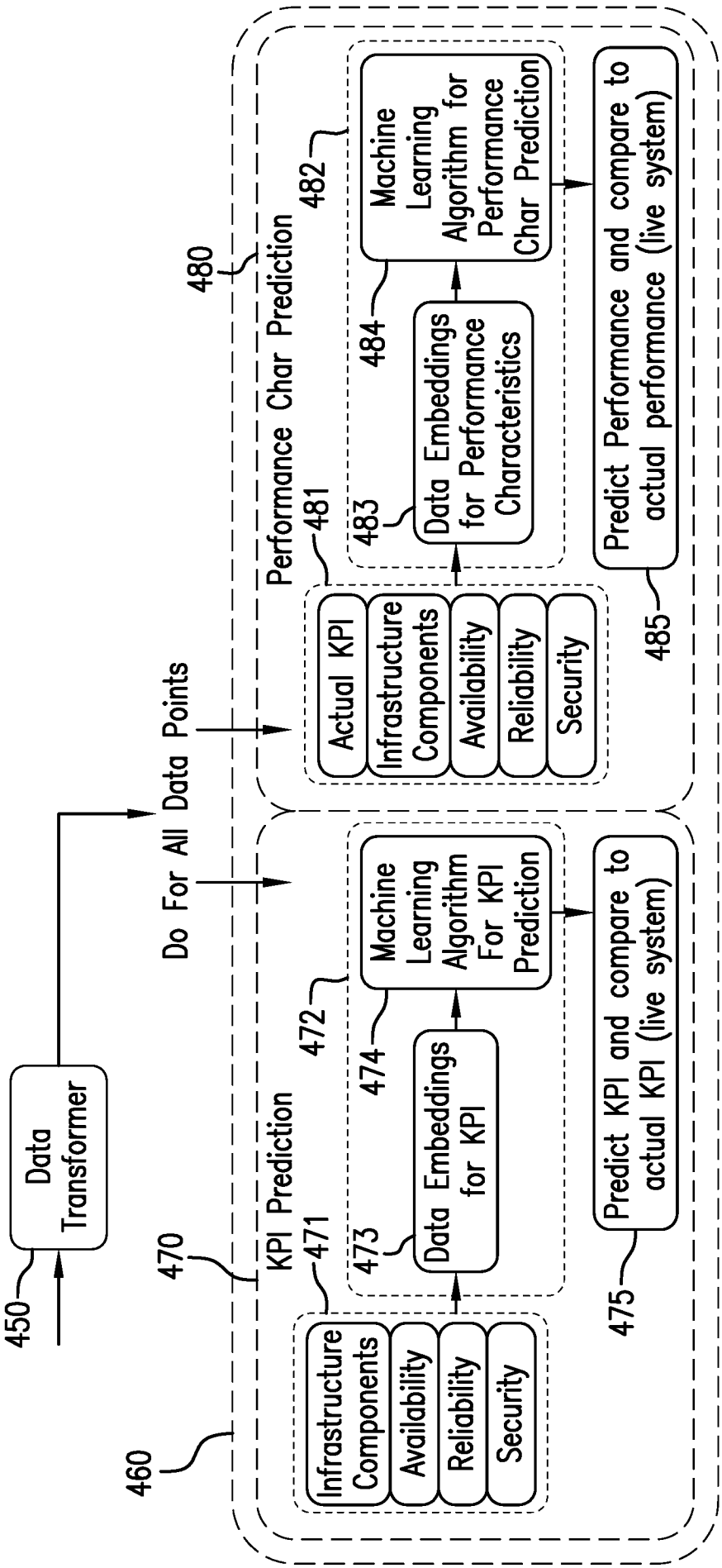
Data Transformer
450
Do For All Data Points
KPI Prediction
470
Infrastructure Components
Availability
Reliability
Security
471
460
Data Embeddings for KPI
473
Machine Learning Algorithm For KPI Prediction
472
474
Predict KPI and compare to actual KPI (live system)
475
Performance Char Prediction
480
Actual KPI
Infrastructure Components
Availability
Reliability
Security
481
Data Embeddings for Performance Characteristics
483
Machine Learning Algorithm for Performance Char Prediction
482
484
Predict Performance and compare to actual performance (live system)
485

FIG.6A

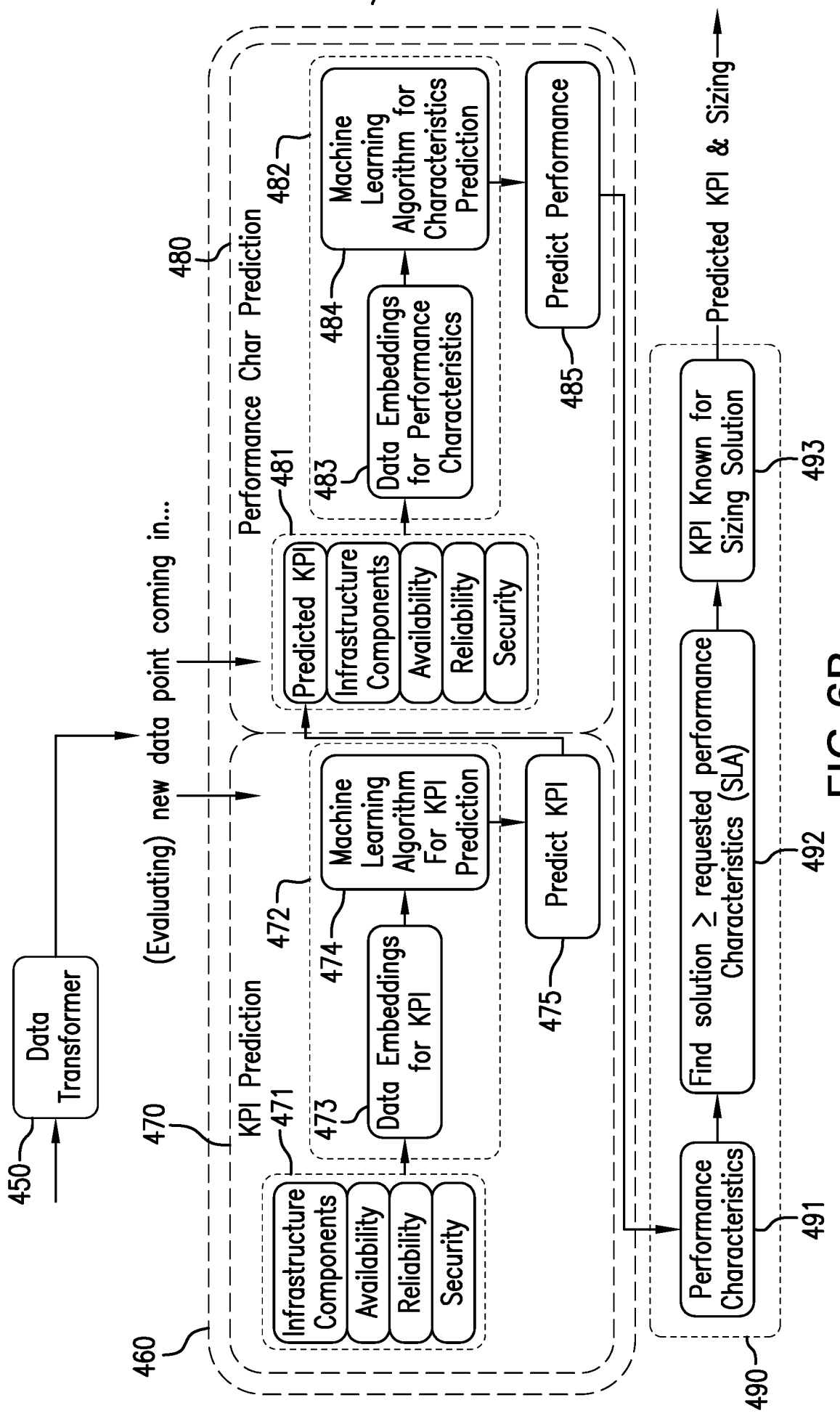
Data Transformer
450
(Evaluating) new data point coming in...
460
KPI Prediction
470
Infrastructure Components
Availability
Reliability
Security
471
Data Embeddings for KPI
473
Machine Learning Algorithm For KPI Prediction
472
474
Predict KPI
475
Performance Char Prediction
480
Predicted KPI
Infrastructure Components
Availability
Reliability
Security
481
Data Embeddings for Performance Characteristics
483
Machine Learning Algorithm for Characteristics Prediction
482
484
Predict Performance
485
Performance Characteristics
491
Find solution ≥ requested performance Characteristics (SLA)
492
KPI Known for Sizing Solution
493
Predicted KPI & Sizing
490

FIG.6B

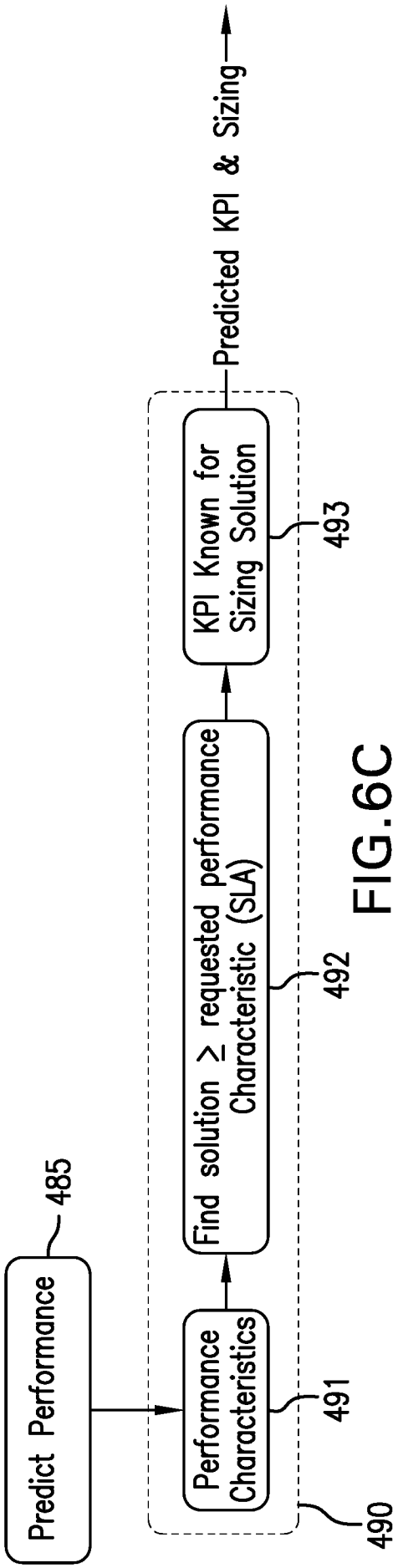
Predict Performance
485
Performance Characteristics
491
Find solution ≥ requested performance Characteristic (SLA)
492
KPI Known for Sizing Solution
493
Predicted KPI & Sizing
490

FIG.6C

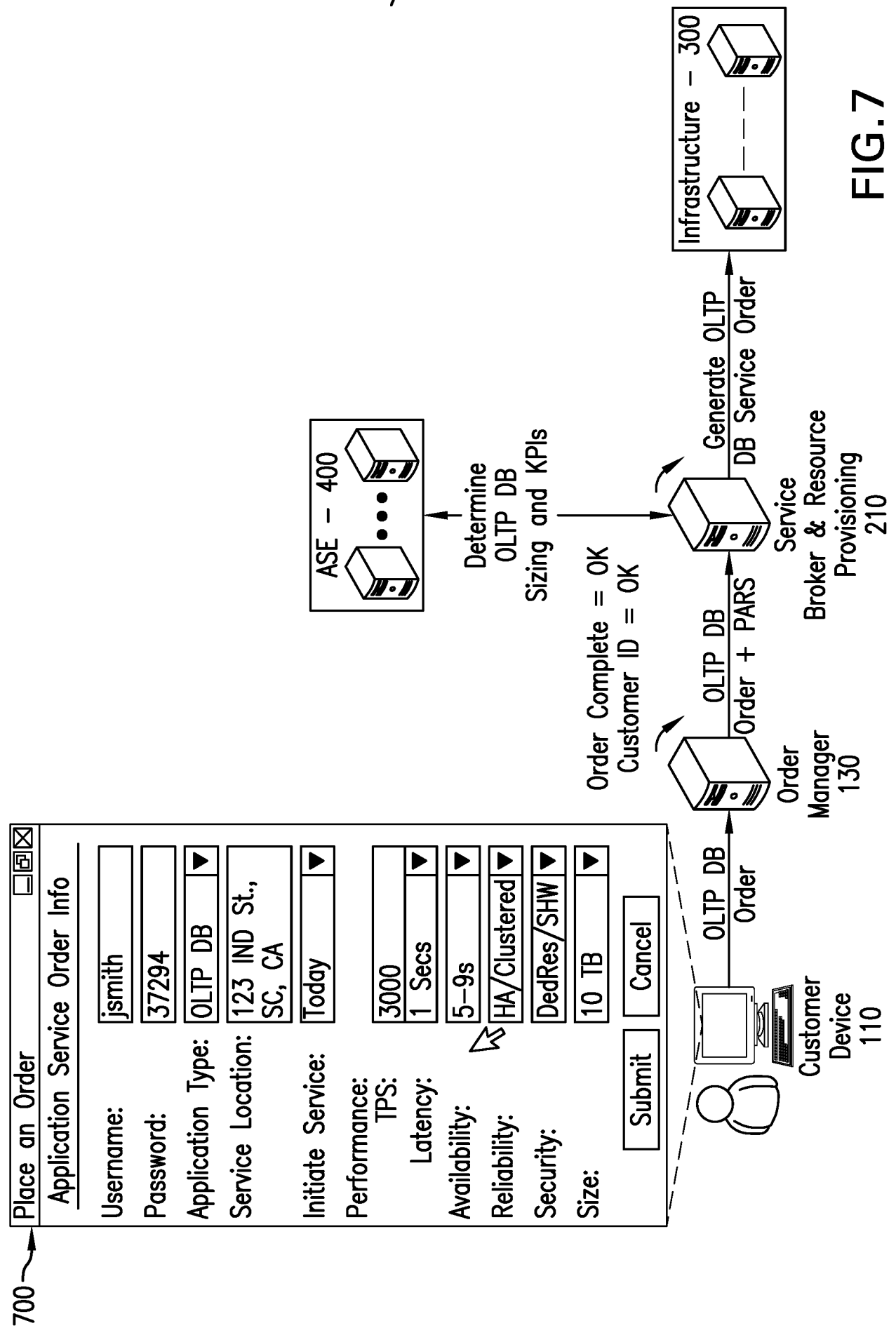
700
Place an Order
Application Service Order Info
Username:
jsmith
Password:
37294
Application Type:
OLTP DB
Service Location:
123 IND St., SC, CA
Initiate Service:
Today
Performance:
TPS:
3000
Latency:
1 Secs
Availability:
5–9s
Reliability:
HA/Clustered
Security:
DedRes/SHW
Size:
10 TB
Submit
Cancel
OLTP DB
Order
Customer Device 110
Order Manager 130
OLTP DB
Order + PARS
Order Complete = OK
Customer ID = OK
Service Broker & Resource Provisioning 210
Generate OLTP
DB Service Order
Infrastructure – 300
Determine OLTP DB Sizing and KPIs
ASE – 400

FIG. 7

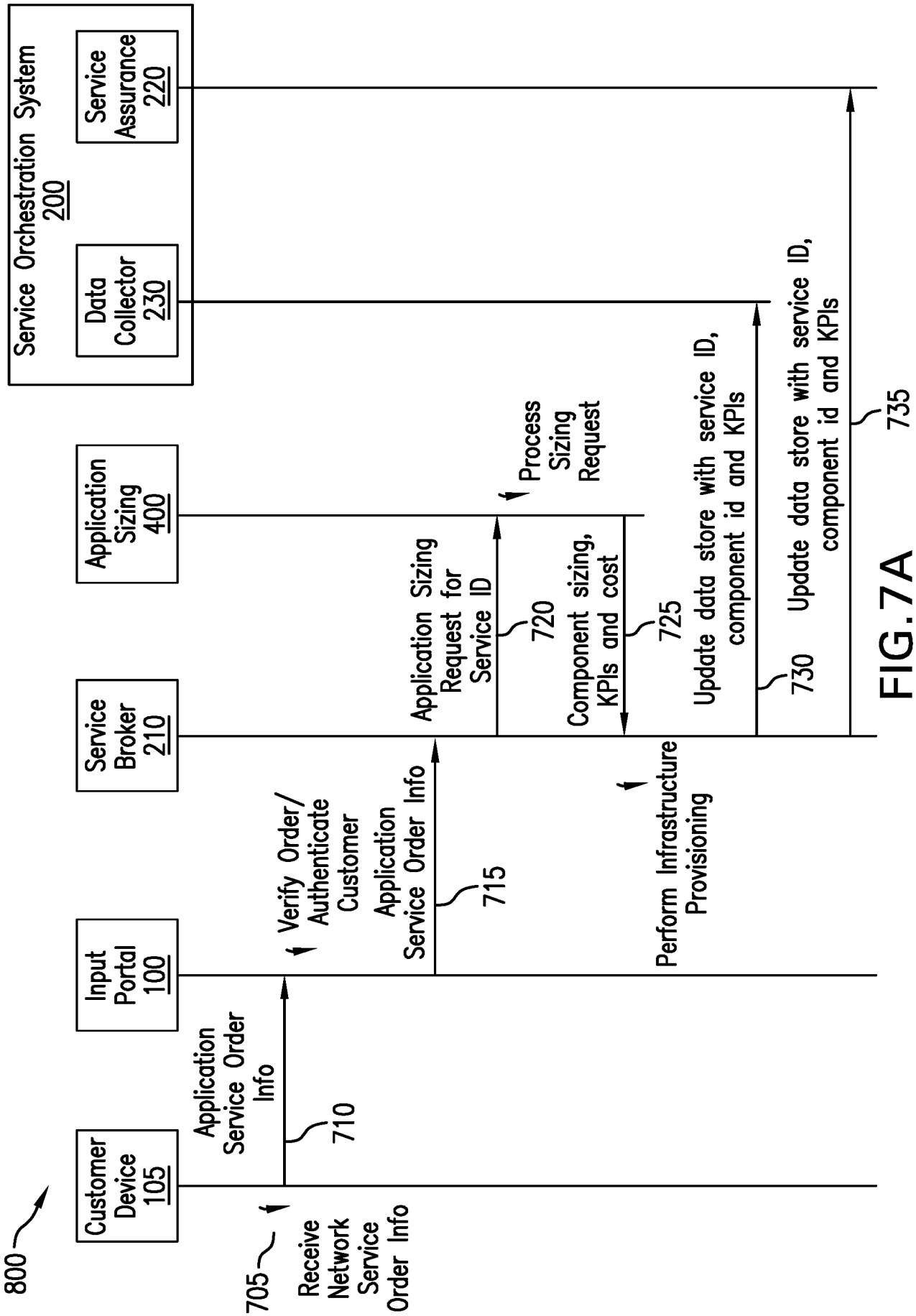
800
Customer Device 105
Input Portal 100
Service Broker 210
Application Sizing 400
Service Orchestration System 200
Data Collector 230
Service Assurance 220
705
Receive Network Service Order Info
Application Service Order Info
710
Verify Order/ Authenticate Customer
Application Service Order Info
715
Application Sizing Request for Service ID
720
Process Sizing Request
Component sizing, KPIs and cost
725
Perform Infrastructure Provisioning
Update data store with service ID, component id and KPIs
730
Update data store with service ID, component id and KPIs
735

FIG. 7A

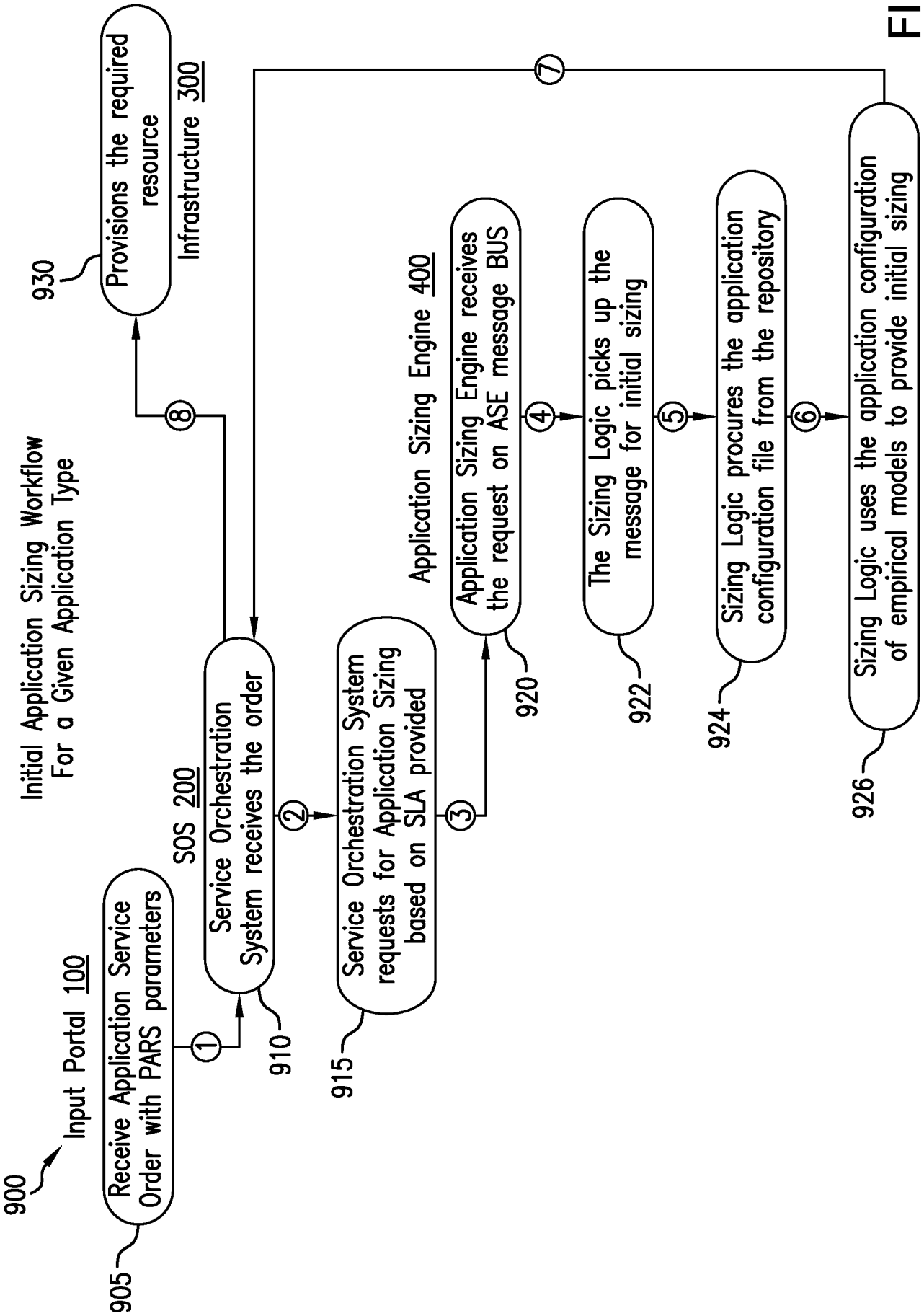
Initial Application Sizing Workflow
For a Given Application Type
900
Input Portal 100
Receive Application Service Order with PARS parameters
905
1
SOS 200
Service Orchestration System receives the order
910
2
Service Orchestration System requests for Application Sizing based on SLA provided
915
3
Application Sizing Engine 400
Application Sizing Engine receives the request on ASE message BUS
920
4
The Sizing Logic picks up the message for initial sizing
922
5
Sizing Logic procures the application configuration file from the repository
924
6
Sizing Logic uses the application configuration of empirical models to provide initial sizing
926
7
8
Provisions the required resource
Infrastructure 300
930

FIG.8

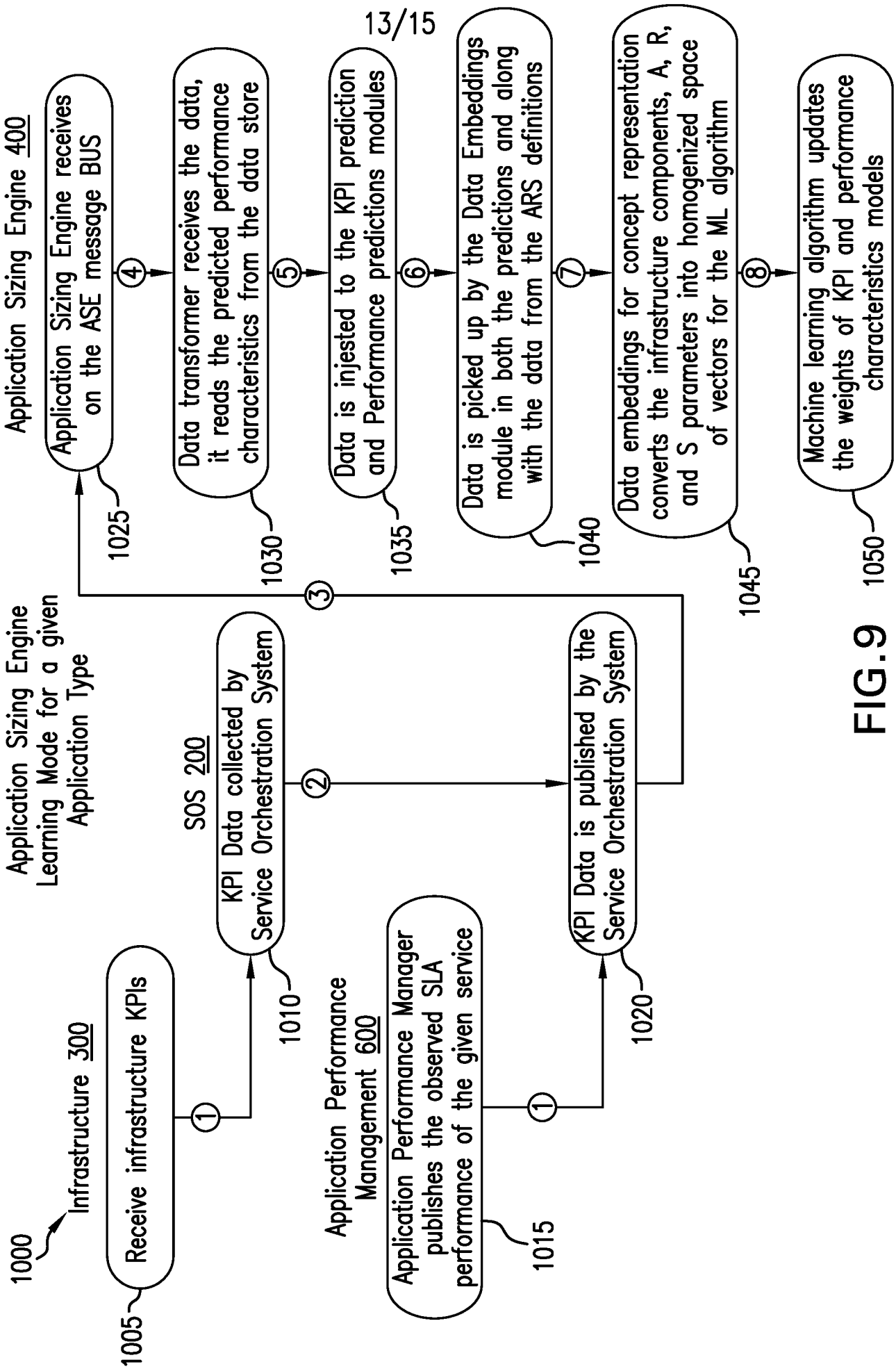
1000
Infrastructure 300
Receive infrastructure KPIs
1005
Application Sizing Engine Learning Mode for a given Application Type
SOS 200
KPI Data collected by Service Orchestration System
1010
Application Performance Management 600
Application Performance Manager publishes the observed SLA performance of the given service
1015
KPI Data is published by the Service Orchestration System
1020
Application Sizing Engine 400
Application Sizing Engine receives on the ASE message BUS
1025
Data transformer receives the data, it reads the predicted performance characteristics from the data store
1030
Data is injested to the KPI prediction and Performance predictions modules
1035
Data is picked up by the Data Embeddings module in both the predictions and along with the data from the ARS definitions
1040
Data embeddings for concept representation converts the infrastructure components, A, R, and S parameters into homogenized space of vectors for the ML algorithm
1045
Machine learning algorithm updates the weights of KPI and performance characteristics models
1050
1
2
3
4
5
6
7
8

FIG.9

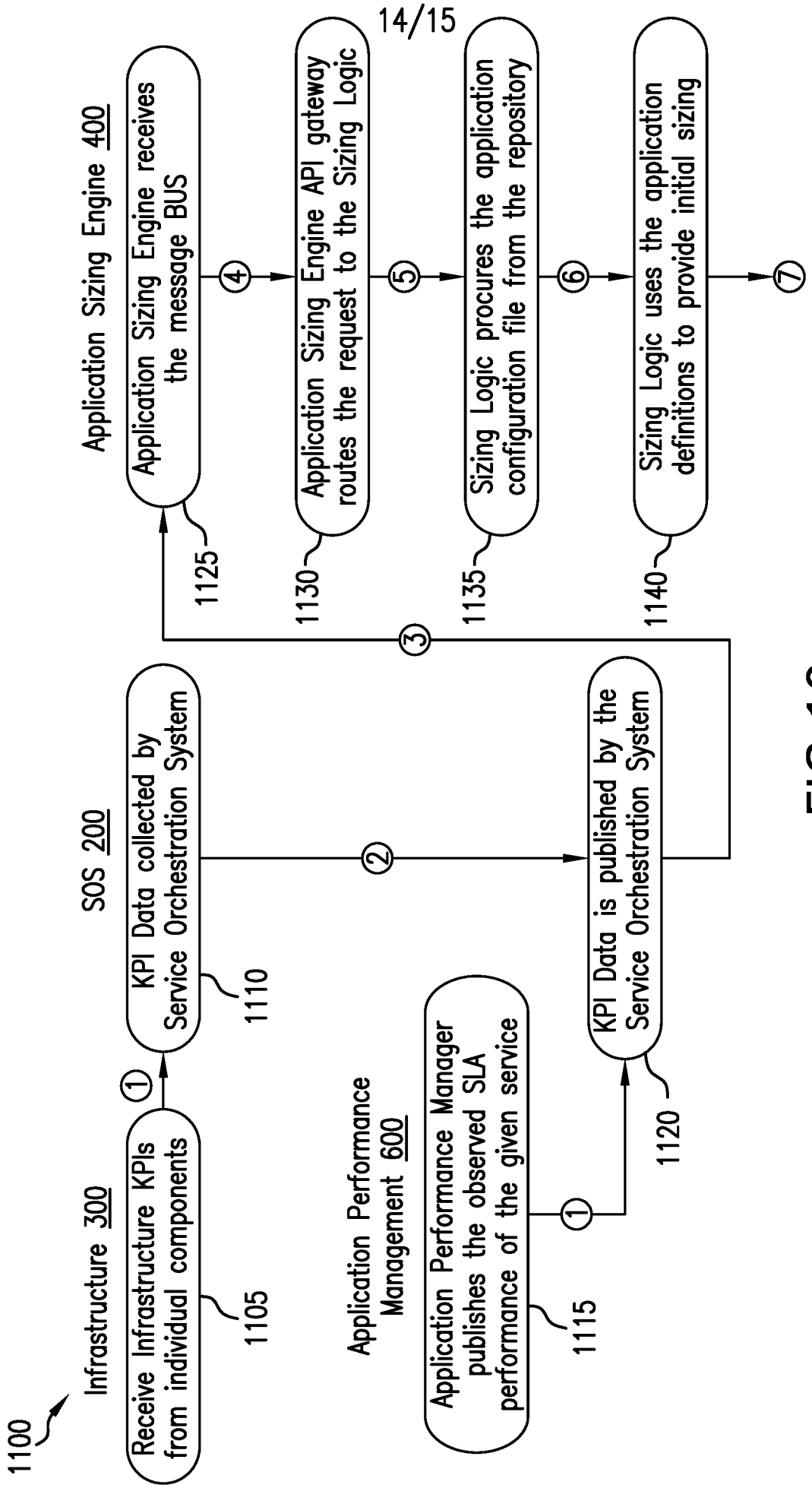
1100
Infrastructure 300
Receive Infrastructure KPIs from individual components
1105
1
SOS 200
KPI Data collected by Service Orchestration System
1110
2
Application Performance Management 600
Application Performance Manager publishes the observed SLA performance of the given service
1115
1
KPI Data is published by the Service Orchestration System
1120
3
Application Sizing Engine 400
Application Sizing Engine receives the message BUS
1125
4
Application Sizing Engine API gateway routes the request to the Sizing Logic
1130
5
Sizing Logic procures the application configuration file from the repository
1135
6
Sizing Logic uses the application definitions to provide initial sizing
1140
7

FIG.10

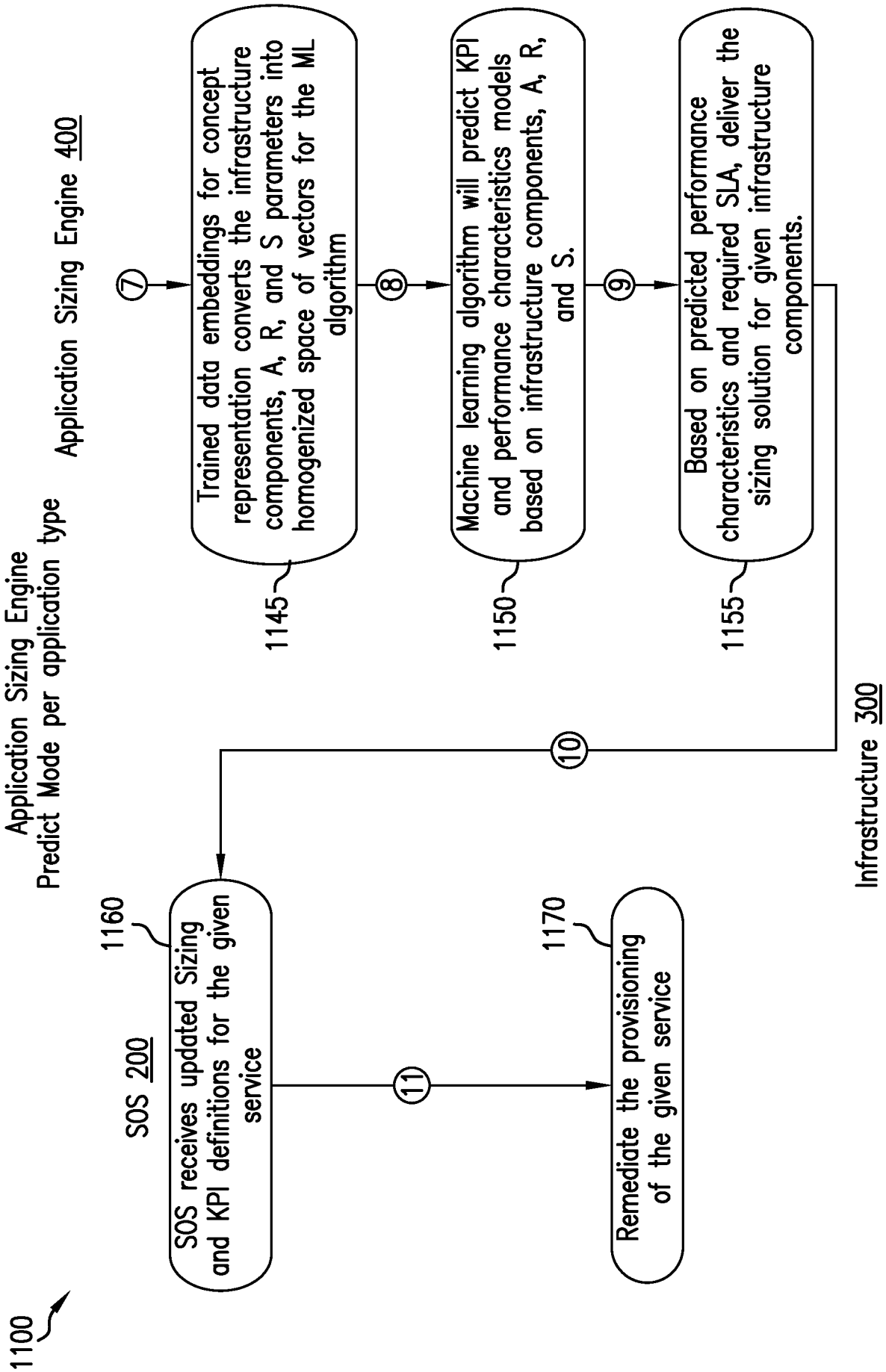
1100
Application Sizing Engine
Predict Mode per application type
Application Sizing Engine 400
SOS 200
1160
SOS receives updated Sizing and KPI definitions for the given service
7
Trained data embeddings for concept representation converts the infrastructure components, A, R, and S parameters into homogenized space of vectors for the ML algorithm
1145
8
Machine learning algorithm will predict KPI and performance characteristics models based on infrastructure components, A, R, and S.
1150
9
Based on predicted performance characteristics and required SLA, deliver the sizing solution for given infrastructure components.
1155
Infrastructure 300
10
11
1170
Remediate the provisioning of the given service

FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/33942

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G06F 9/44, H04L 12/24 (2021.01)

CPC - G06F 8/77, H04L 41/5054, H04L 67/10, G06F 8/61, G06F 8/20, G06F 8/70, G06F 9/45558

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)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2019/0265971 A1 (C3 IoT, Inc.) 29 August 2019 (29.08.2019) entire document (especially Figs. 1-41 & para [0141], [0170], [0223], [0246], [0310], [0437], [0488], [0505], [0310], [0516], [0517], [0539], [0585], [0592], [0593], [0603], [0606], [0616], [0627], [0633]).	1-20
Y	US 2020/0050494 A1 (Intel Corporation) 13 February 2020 (13.02.2020) entire document (especially para Figs. 1-31 & para [0104], [0105], [0109], [0111], [0124], [0130], [0131], [0148], [0176], [0207], [0272]).	1-20
A	US 2019/0102700 A1 (Oracle International Corporation) 04 April 2019 (04.04.2019) entire document.	1-20
A	US 2009/0112932 A1 (Skierkowski et al.) 30 April 2009 (30.04.2009) entire document.	1-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"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)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"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
"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
"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
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 July 2021 (28.07.2021)	AUG 25 2021
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Kari Rodriquez Telephone No. PCT Helpdesk: 571-272-4300