US 20230232195A1

(19) **United States**

(12) **Patent Application Publication**

Biswas et al.

(10) **Pub. No.: US 2023/0232195 A1**

(43) **Pub. Date: Jul. 20, 2023**

(54) **COLLECTIVE SCALING OF APPLICATIONS**

(71) Applicant: **VMware, Inc.**, Palo Alto, CA (US)

(72) Inventors: **Sudipta Biswas**, Bangalore (IN); **Monotosh Das**, Bangalore (IN); **Hemant Kumar Shaw**, Bangalore (IN); **Shubham Chauhan**, Faridabad (IN)

(21) Appl. No.: **17/729,776**

(22) Filed: **Apr. 26, 2022**

(30) **Foreign Application Priority Data**

Jan. 19, 2022	(IN)	202241003041
Jan. 19, 2022	(IN)	202241003045

Publication Classification

(51) **Int. Cl.**

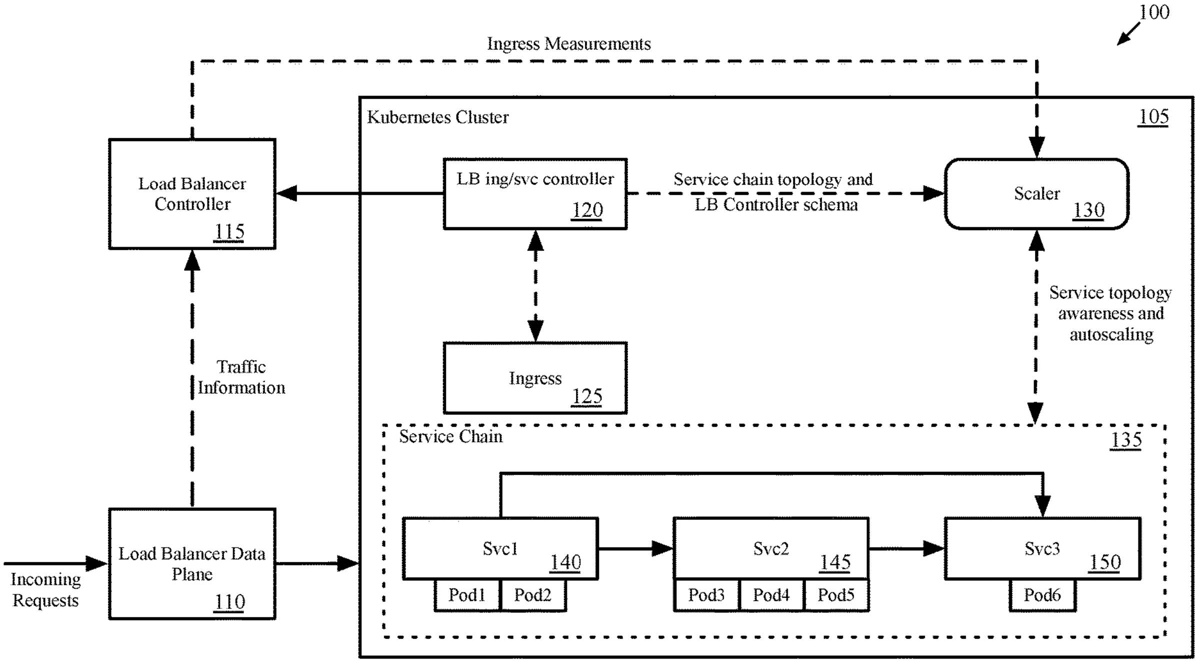
H04W 4/50	(2006.01)
H04W 28/02	(2006.01)
H04L 41/40	(2006.01)

(52) **U.S. Cl.**

CPC ***H04W 4/50*** (2018.02); ***H04W 28/0289*** (2013.01); ***H04L 41/40*** (2022.05)

(57) **ABSTRACT**

Some embodiments provide a method for scaling a service chain that includes multiple services, each of which is provided by one or more instances of the service. The method identifies that a first service in the service chain has received a number of requests. For each service in the service chain, the method (i) identifies a scaling factor that estimates a portion of requests received at the first service that will be subsequently received at the service and (ii) deploys a number of additional instances of the service based on the identified scaling factor for the service and the number of requests received at the first service.


100
Ingress Measurements
Kubernetes Cluster
105
Load Balancer Controller
115
LB ing/svc controller
120
Service chain topology and LB Controller schema
Scaler
130
Service topology awareness and autoscaling
Ingress
125
Traffic Information
Service Chain
135
Svc1
140
Svc2
145
Svc3
150
Pod1
Pod2
Pod3
Pod4
Pod5
Pod6
Incoming Requests
Load Balancer Data Plane
110