

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 1 月 13 日 (13.01.2022)



(10) 国际公布号
WO 2022/007406 A1

(51) 国际专利分类号:
H04L 12/24 (2006.01)

(21) 国际申请号: PCT/CN2021/076995

(22) 国际申请日: 2021 年 2 月 20 日 (20.02.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010664169.2 2020 年 7 月 10 日 (10.07.2020) CN

(71) 申请人: 苏州浪潮智能科技有限公司
(**INSPUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.**) [CN/CN]; 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。

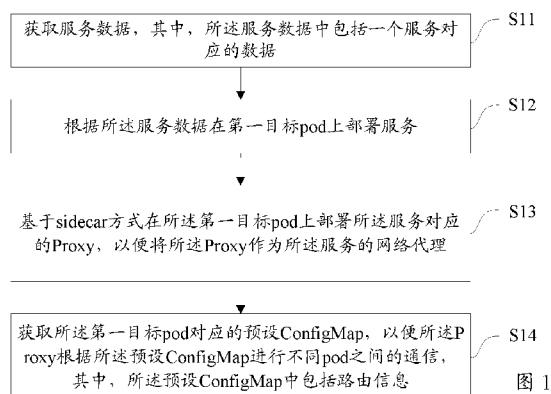
(72) 发明人: 陈清山 (**CHEN, Qingshan**); 中国江苏省苏州市吴中区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。

(74) 代理人: 北京集佳知识产权代理有限公司
(**UNITALEN ATTORNEYS AT LAW**); 中国北京市朝阳区建国门外大街 22 号赛特广场 7 层, Beijing 100004 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): 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,

(54) **Title:** K8S-BASED SERVICE DEPLOYMENT METHOD, APPARATUS, DEVICE, AND MEDIUM

(54) 发明名称: 一种基于k8s的服务部署方法、装置、设备、介质



- S11 Acquire service data, the service data comprising data corresponding to one service
S12 Deploy a service on a first target pod according to the service data
S13 Deploy, on the basis of a sidecar means, a proxy corresponding to the service on the first target pod, so as to use the proxy as a network proxy of the service
S14 Acquire a preset ConfigMap corresponding to the first target pod, so that the proxy performs communication between different pods according to the preset ConfigMap, the preset ConfigMap comprising routing information

(57) **Abstract:** A k8s-based service deployment method, apparatus, device, and medium. The method comprises: acquiring service data, the service data comprising data corresponding to one service; deploying a service on a first target pod according to the service data; deploying, on the basis of a sidecar means, a proxy corresponding to the service on the first target pod, so as to use the proxy as a network proxy of the service; and acquiring a preset ConfigMap corresponding to the first target pod, so that the proxy performs communication between different pods according to the preset ConfigMap, the preset ConfigMap comprising routing information. As

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) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

such, only one service is deployed by one pod, and a separate service does not need to be deployed to string together each independent service on one pod, thus reducing the workload in a service deployment process, and improving resource utilization. A plug-in may be added to the service by means of the proxy, and after the plug-in is added, compilation of the entire pod does not need to be reperformed.

(57) 摘要: 一种基于k8s的服务部署方法、装置、设备、介质, 该方法包括: 获取服务数据, 其中, 所述服务数据中包括一个服务对应的数据; 根据所述服务数据在第一目标pod上部署服务; 基于sidecar方式在所述第一目标pod上部署所述服务对应的Proxy, 以便将所述Proxy作为所述服务的网络代理; 获取所述第一目标pod对应的预设ConfigMap, 以便所述Proxy根据所述预设ConfigMap进行不同pod之间的通信, 其中, 所述预设ConfigMap中包括路由信息。这样一个pod只部署一个服务, 不需要再部署一个单独的服务将一个pod上的各个独立服务串起来, 减少了服务部署过程中的工作量, 且提高了资源利用率, 可以通过Proxy为服务增加插件, 增加插件后不需要重新进行整个pod的编译。

一种基于 k8s 的服务部署方法、装置、设备、介质

本申请要求于 2020 年 07 月 10 日提交中国专利局、申请号为 CN202010664169.2、发明名称为“一种基于 k8s 的服务部署方法、装置、设备、介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及计算机技术领域，特别涉及一种基于 k8s 的服务部署方法、装置、设备、介质。

背景技术

- 10 在k8s（也即kubernetes）中通常将多个独立模型部署在同一个pod上，并开发一个独立服务来调用这个pod上的独立模型，也即这个独立服务负责按照设计的顺序依次调用pod中的独立模型。同时将上一个pod的输出作为当前pod的输入。其中，一个独立模型对应一个服务。然而在这个模式下，需要伸缩时，只能对整个pod进行伸缩，减低了伸缩过程中的灵活性。此外，
- 15 需要单独开发一个独立服务，用来将pod中的独立模型串联起来，增加了服务开发过程中的工作量。且在修改某一个独立模型的时候，如调整顺序，或者增加、删除某一独立模型时，需要重新编译整个pod，这样会导致修改服务需要耗费的时间较长。且在资源占用上，部署多个独立模型的pod无法达到资源利用率最高，类似木桶原理，瓶颈会存在于最消耗资源的那个模
- 20 型上。

发明内容

- 有鉴于此，本申请的目的在于提供一种基于k8s的服务部署方法、装置、设备、介质，能够提高服务部署的灵活性，且提高了资源利用率。其具体
- 25 方案如下：

第一方面，本申请公开了一种基于 k8s 的服务部署方法，包括：
获取服务数据，其中，所述服务数据中包括一个服务对应的数据；
根据所述服务数据在第一目标 pod 上部署服务；

-2-

基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理；

获取所述第一目标 pod 对应的预设 ConfigMap，以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信，其中，所述预设 ConfigMap 5 中包括路由信息。

可选地，所述基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy 之后，还包括：

基于 TCP 协议建立所述服务和所述 Proxy 之间的通信连接。

可选地，所述基于 k8s 的服务部署方法，还包括：

10 获取所述第一目标 pod 对应的伸缩规则；

对所述伸缩规则进行存储，以便根据系统的运行负载和所述伸缩规则对所述第一目标 pod 进行伸缩。

可选地，所述获取所述第一目标 pod 对应的预设 ConfigMap 之后，还包括：

15 当所述 Proxy 接收到待传输数据时，通过所述服务对所述待传输数据进行处理，得到处理后待传输数据；

通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod，其中，所述第二目标 pod 为根据所述预设 ConfigMap 中的路由信息确定出的。

20 可选地，所述通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod 之前，还包括：

基于 HTTP 协议建立所述第一目标 pod 中的 Proxy 与所述第二目标 pod 中的 Proxy 之间的通信连接。

可选地，所述通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod 之前，还包括：

25 根据所述第一目标 pod 中的 Proxy 名称确定所述第一目标 pod 在所述预设 ConfigMap 中的路由信息中所处的当前节点；

根据所述路由信息和所述第一目标 pod 所处的当前节点确定出所述第二目标 pod。

可选地，所述根据所述路由信息和所述第一目标 pod 所处的当前节点

—3—

确定出所述第二目标 pod，包括：

如果所述第一目标 pod 所处的当前节点为所述路由信息中的最后一个节点，则将所述路由信息中的第一个节点确定为所述第二目标 pod。

第二方面，本申请公开了一种基于 k8s 的服务部署装置，包括：

5 第一数据获取模块，用于获取服务数据，其中，所述服务数据中包括一个服务对应的数据；

第一部署模块，用于根据所述服务数据在第一目标 pod 上部署服务；

第二部署模块，用于基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理；

10 第二数据获取模块，用于获取所述第一目标 pod 对应的预设 ConfigMap，以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信，其中，所述预设 ConfigMap 中包括路由信息。

第三方面，本申请公开了一种电子设备，包括：

存储器和处理器；

15 其中，所述存储器，用于存储计算机程序；

所述处理器，用于执行所述计算机程序，以实现前述公开的基于 k8s 的服务部署方法。

20 第四方面，本申请公开了一种计算机可读存储介质，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现前述公开的基于 k8s 的服务部署方法。

可见，本申请先获取服务数据，其中，所述服务数据中包括一个服务对应的数据，然后根据所述服务数据在第一目标 pod 上部署服务，并基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理，再获取所述第一目标 pod 对应的预设
25 ConfigMap，以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信，其中，所述预设 ConfigMap 中包括路由信息。由此可见，本申请中在获取包括一个服务对应的数据之后，根据接收到的服务数据将这个服务部署在一个 pod 上，这样一个 pod 只部署一个服务，便不需要再部署一个单独的服务来将一个 pod 上的各个独立服务串起来，减少了服务部署过

—4—

程中的工作量，且可以只根据 pod 上这个服务自身的负载对这个 pod 进行伸缩，提高了资源利用率，此外还可以通过 Proxy 为服务增加插件，增加插件后不需要重新进行整个 pod 的编译，且 Proxy 可以根据预设 ConfigMap 中的路由信息进行不同 pod 之间的通信，使得 pod 之间通信更方便。

5

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

10

图1为本申请公开的一种基于k8s的服务部署方法流程图；

图2为本申请公开的一种具体的基于k8s的服务部署方法流程图；

图3为本申请公开的一种Proxy和服务之间的通信示意图；

图4为本申请公开的一种预设ConfigMap示意图；

15

图5为本申请公开的一种pod部署示意图；

图6为本申请公开的一种pod伸缩示意图；

图7为本申请公开的一种不同Proxy之间的通信示意图；

图8为本申请公开的一种基于k8s的服务部署装置结构示意图；

图9为本申请公开的一种电子设备结构图。

20

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的

25

范围。

参见图1所示，本申请实施例公开了一种基于k8s的服务部署方法，该方法包括：

—5—

步骤 S11: 获取服务数据, 其中, 所述服务数据中包括一个服务对应的数据。

在具体的实施过程中, 需要先获取服务数据, 其中, 所述服务数据中包括一个服务对应的数据, 以便根据所述服务数据在第一目标 pod 上部署
5 相应的服务。

步骤 S12: 根据所述服务数据在第一目标 pod 上部署服务。

相应地, 在获取到所述服务数据之后, 便可以根据所述服务数据在第一目标 pod 上部署服务, 其中, 所述第一目标 pod 可以是用户选择的 pod。

步骤 S13: 基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的
10 的 Proxy, 以便将所述 Proxy 作为所述服务的网络代理。

在所述第一目标 pod 上部署所述服务之后, 还需要在所述第一目标 pod 上部署所述服务对应的 Proxy, 以便将所述 Proxy 作为所述服务的网络代理。具体的, 可以基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy。sidecar 是 k8s 中的一种部署形式, 通过 sidecar 方式, 可以自
15 动为某一类服务增加一个插件, 达到实现某种功能的目的。proxy 采用 sidecar 方式部署, 对于使用者来说是透明的, 也即实现了 Proxy 透明化, Proxy 透明化指的是 Proxy 组件对用户透明, 因为 Proxy 采用 sidecar 方式实现, 会自动为用户的服务增加插件, 所以用户不需要配置 proxy 相关内容, 达到对用户透明。

20 步骤 S14: 获取所述第一目标 pod 对应的预设 ConfigMap, 以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信, 其中, 所述预设 ConfigMap 中包括路由信息。

在所述第一目标 pod 上部署所述 Proxy 之后, 还需要获取所述第一目标 pod 对应的预设 ConfigMap, 以便所述 Proxy 根据所述预设 ConfigMap
25 进行不同 pod 之间的通信, 其中, 所述预设 ConfigMap 中包括路由信息。这样可以利用 Proxy 实现不同 pod 之间的通信。

可见, 本申请先获取服务数据, 其中, 所述服务数据中包括一个服务对应的数据, 然后根据所述服务数据在第一目标 pod 上部署服务, 并基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy, 以便将所述

—6—

Proxy作为所述服务的网络代理，再获取所述第一目标pod对应的预设ConfigMap，以便所述Proxy根据所述预设ConfigMap进行不同pod之间的通信，其中，所述预设ConfigMap中包括路由信息。由此可见，本申请中在获取包括一个服务对应的数据之后，根据接收到的服务数据将这个服务部署在一个pod上，这样一个pod只部署一个服务，便不需要再部署一个单独的服务来将一个pod上的各个独立服务串起来，减少了服务部署过程中的工作量，且可以只根据pod上这个服务自身的负载对这个pod进行伸缩，提高了资源利用率，此外还可以通过Proxy为服务增加插件，增加插件后不需要重新进行整个pod的编译，且Proxy可以根据预设ConfigMap中的路由信息进行不同pod之间的通信，使得pod之间通信更方便。

参见图2所示，本申请实施例公开了一种具体的基于k8s的服务部署方法，该方法包括：

步骤 S21：获取服务数据，其中，所述服务数据中包括一个服务对应的数据。

步骤 S22：根据所述服务数据在第一目标 pod 上部署服务。

步骤S23：基于sidecar方式在所述第一目标pod上部署所述服务对应的Proxy，以便将所述Proxy作为所述服务的网络代理。

步骤S24：基于TCP协议建立所述服务和所述Proxy之间的通信连接。

在所述第一目标 pod 上部署所述 Proxy 之后，还需要基于 TCP（Transmission Control Protocol，传输控制协议）协议建立所述服务和所述Proxy之间的通信连接。具体的，就是Proxy容器对外端口默认为80，通过与服务的port端口建立TCP通信，进行请求和响应的传递。参见图3所示，为Proxy和服务之间的通信原理示意图。Proxy对外端口默认为80，通过与服务的port端口建立TCP通信，进行请求和响应的传递。

步骤S25：获取所述第一目标pod对应的预设ConfigMap，以便所述Proxy根据所述预设ConfigMap进行不同pod之间的通信，其中，所述预设ConfigMap中包括路由信息。

可以理解的是，在建立所述服务和所述Proxy之间的通信连接之后，还

-7-

需要获取所述第一目标pod对应的预设ConfigMap,以便所述Proxy根据所述预设ConfigMap进行不同pod之间的通信,其中,所述预设ConfigMap中包括路由信息。参见图4所示,为预设ConfigMap包括的信息。所述预设ConfigMap包括路由信息,也即图中的RouteMap,路由信息中包括各个节点的名称(name),各个节点的类型(type),所述各个节点的类型包括开始、过程、结束。还包括各个节点的地址,以及当前节点的下一个节点。例如,图4中的路由信息包括3个节点,第一个节点名称Proxy A,节点类型为开始(START),地址为http://ip:port/methodA,下一个节点为Proxy B;第二个节点名称Proxy B,节点类型为过程(PROCESS),地址为http://ip:port/methodB,下一个节点为Proxy C;第三个节点名称Proxy C,节点类型为结束(END),地址为http://ip:port/methodC,下一个节点为空(NULL)。

步骤 S26: 获取所述第一目标 pod 对应的伸缩规则。

在实际应用中,还需要设置所述第一目标 pod 的伸缩规则,以便根据所述伸缩规则对所述第一目标 pod 进行伸缩,以控制所述第一目标 pod 的副本数量。

步骤S27: 对所述伸缩规则进行存储,以便根据系统的运行负载和所述伸缩规则对所述第一目标pod进行伸缩。

在获取到所述伸缩规则之后,便可以对所述伸缩规则进行存储,以便根据系统的运行负载和所述伸缩规则对所述第一目标pod进行伸缩。也即,在对所述伸缩规则进行存储之后,便可以根据所述第一目标pod所处的系统的负载情况以及所述伸缩规则,对所述第一目标pod的副本数量进行控制。例如,在所述系统的负载较小时,可以根据所述伸缩规则增加所述第一目标pod的副本数量,当在所述系统的负载较大时,可以根据所述伸缩规则减少所述第一目标pod的副本数量。

参见图5所示,为pod部署示意图。根据实际应用场景可以在多个pod上部署服务(service)以及各个服务对应的Proxy,例如,在pod A上部署service A以及service A对应的Proxy A,在pod B上部署service B以及service B对应的Proxy B,在pod C上部署service C以及service C对应的Proxy C,在

—8—

pod D上部署service D以及service D对应的Proxy D, 在pod E上部署service E以及service E对应的Proxy E。Proxy A可以与Proxy B进行通信, Proxy B可以与Proxy C、Proxy D进行通信, Proxy C和Proxy D均可以和Proxy E进行通信。

- 5 参见图6所示, 为pod伸缩示意图。根据所述系统的负载和各个pod的伸缩, 控制各个pod的副本数量, 各个pod的副本数量可以不一样, 例如, 根据系统的负载和pod A的伸缩规则, 控制pod A的副本数量为1, 根据系统的负载和pod B的伸缩规则, 控制pod B的副本数量为0, 根据系统的负载和pod C的伸缩规则, 控制pod C的副本数量为2, 根据系统的负载和pod D的伸缩规则, 控制pod D的副本数量为0, 根据系统的负载和pod E的伸缩规则, 控制pod E的副本数量为0。

步骤 S28: 当所述 Proxy 接收到待传输数据时, 通过所述服务对所述待传输数据进行处理, 得到处理后待传输数据。

- 15 在pod部署完毕后, 当所述第一目标pod中的Proxy接收到待传输数据时, 通过所述服务对所述待传输数据进行处理, 得到处理后待传输数据。具体的, 就是当所述第一目标pod中的Proxy接收到待传输数据, 将所述待传输数据发送到所述第一目标pod的服务中, 所述第一目标pod中的服务对所述待传输数据进行处理, 得到所述处理后待传输数据。

- 20 步骤S29: 通过所述Proxy将所述处理后待传输数据发送给第二目标pod, 其中, 所述第二目标pod为根据所述预设ConfigMap中的路由信息确定出的。

得到所述处理后待传输数据之后, 便可以通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod, 其中, 所述第二目标 pod 为根据所述预设 ConfigMap 中的路由信息确定出的。

- 25 所述通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod 之前, 还包括: 根据所述第一目标 pod 中的 Proxy 名称确定所述第一目标 pod 在所述预设 ConfigMap 中的路由信息中所处的当前节点; 根据所述路由信息和所述第一目标 pod 所处的当前节点确定出所述第二目标 pod。确定出所述第二目标 pod 之后, 便可以从所述路由信息中确定出所述第二目标 pod

的地址信息，便可以将所述处理后待传输数据发送到所述第二目标 pod。

在根据所述路由信息和所述第一目标 pod 所处的当前节点确定出所述第二目标 pod 的过程中，包括：如果所述第一目标 pod 所处的当前节点为所述路由信息中的最后一个节点，则将所述路由信息中的第一个节点确定
5 为所述第二目标 pod。也即，当接收到待传输数据的 Proxy 所述路由信息中的最后一个 Proxy 时，该 Proxy 会将接收到的处理后待传输数据发送到所述路由信息中的第一个 Proxy，以便第一个 Proxy 将最终结果返回给调用端。

在实施过程中，所述通过所述 Proxy 将所述处理后待传输数据发送给
10 第二目标 pod 之前，还包括：基于 HTTP（HyperText Transfer Protocol，超文本传输协议）协议建立所述第一目标 pod 中的 Proxy 与所述第二目标 pod 中的 Proxy 之间的通信连接。参见图 7 所示，为不同 Proxy 之间的通信示意图。不同的 Proxy 都对外暴露端口 80，当前 Proxy 与另一个 Proxy 的 80 端口建立 HTTP 通信，进行请求和响应的传递。

15

参见图8所示，本申请实施例公开了一种基于k8s的服务部署装置，包括：

第一数据获取模块 11，用于获取服务数据，其中，所述服务数据中包括一个服务对应的数据；

20 第一部署模块 12，用于根据所述服务数据在第一目标 pod 上部署服务；

第二部署模块 13，用于基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理；

第二数据获取模块 14，用于获取所述第一目标 pod 对应的预设 ConfigMap，以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信，其中，所述预设 ConfigMap 中包括路由信息。
25

可见，本申请先获取服务数据，其中，所述服务数据中包括一个服务对应的数据，然后根据所述服务数据在第一目标 pod 上部署服务，并基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理，再获取所述第一目标 pod 对应的预设

ConfigMap, 以便所述Proxy根据所述预设ConfigMap进行不同pod之间的通信, 其中, 所述预设ConfigMap中包括路由信息。由此可见, 本申请中在获取包括一个服务对应的数据之后, 根据接收到的服务数据将这个服务部署在一个pod上, 这样一个pod只部署一个服务, 便不需要再部署一个单独的服务来将一个pod上的各个独立服务串起来, 减少了服务部署过程中的工作量, 且可以只根据pod上这个服务自身的负载对这个pod进行伸缩, 提高了资源利用率, 此外还可以通过Proxy为服务增加插件, 增加插件后不需要重新进行整个pod的编译, 且Proxy可以根据预设ConfigMap中的路由信息进行不同pod之间的通信, 使得到pod之间通信更方便。

5 进一步的, 所述基于k8s的服务部署装置, 还包括:

第一通信建立模块, 用于基于TCP协议建立所述服务和所述Proxy之间的通信连接。

进一步的, 所述基于k8s的服务部署装置, 还包括:

第三数据获取模块, 用于获取所述第一目标 pod 对应的伸缩规则;

15 数据存储模块, 用于对所述伸缩规则进行存储, 以便根据系统的运行负载和所述伸缩规则对所述第一目标pod进行伸缩。

具体的, 所述第一目标 pod 中的服务, 用于当所述 Proxy 接收到待传输数据时, 对所述待传输数据进行处理, 得到处理后待传输数据;

相应的, 所述第一目标pod中的Proxy, 用于将所述处理后待传输数据
20 发送给第二目标pod, 其中, 所述第二目标pod为根据所述预设ConfigMap中的路由信息确定出的。

进一步的, 所述基于k8s的服务部署装置, 还包括:

第二通信建立模块, 用于基于HTTP协议建立所述第一目标pod中的Proxy与所述第二目标pod中的Proxy之间的通信连接。

25 进一步的, 所述基于 k8s 的服务部署装置, 还包括:

第一节点确定模块, 用于根据所述第一目标 pod 中的 Proxy 名称确定所述第一目标 pod 在所述预设 ConfigMap 中的路由信息中所处的当前节点;

第二节点确定模块, 用于根据所述路由信息和所述第一目标pod所处的

当前节点确定出所述第二目标pod。

具体的，所述第二节点确定模块，具体用于：

如果所述第一目标pod所处的当前节点为所述路由信息中的最后一个节点，则将所述路由信息中的第一个节点确定为所述第二目标pod。

5

进一步的，参见图9所示，本申请实施例还公开了一种电子设备，包括：处理器21和存储器22。

其中，所述存储器22，用于存储计算机程序；所述处理器21，用于执行所述计算机程序，以实现前述实施例中公开的基于k8s的服务部署方法。

10 其中，关于上述基于k8s的服务部署方法的具体过程可以参考前述实施例中公开的相应内容，在此不再进行赘述。

进一步的，本申请实施例还公开了一种计算机可读存储介质，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现前述任一实施
15 例中公开的基于k8s的服务部署方法。

其中，关于上述基于k8s的服务部署方法的具体过程可以参考前述实施例中公开的相应内容，在此不再进行赘述。

本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似部分互相参见
20 即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于
25 随机存储器（RAM）、内存、只读存储器（ROM）、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

最后，还需要说明的是，在本文中，诸如第一和第二之类的关系术语仅仅用来将一个实体或者操作与另一个实体或者操作区分开来，而不一定

要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得一系列包含其他要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

10 以上对本申请所提供的一种基于k8s的服务部署方法、装置、设备、介质进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求

1、一种基于 k8s 的服务部署方法，其特征在于，包括：

获取服务数据，其中，所述服务数据中包括一个服务对应的数据；

根据所述服务数据在第一目标 pod 上部署服务；

5 基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy，以便将所述 Proxy 作为所述服务的网络代理；

获取所述第一目标 pod 对应的预设 ConfigMap，以便所述 Proxy 根据所述预设 ConfigMap 进行不同 pod 之间的通信，其中，所述预设 ConfigMap 中包括路由信息。

10 2、根据权利要求 1 所述的基于 k8s 的服务部署方法，其特征在于，所述基于 sidecar 方式在所述第一目标 pod 上部署所述服务对应的 Proxy 之后，还包括：

基于 TCP 协议建立所述服务和所述 Proxy 之间的通信连接。

15 3、根据权利要求 1 所述的基于 k8s 的服务部署方法，其特征在于，还包括：

获取所述第一目标 pod 对应的伸缩规则；

对所述伸缩规则进行存储，以便根据系统的运行负载和所述伸缩规则对所述第一目标 pod 进行伸缩。

20 4、根据权利要求 1 至 3 任一项所述的基于 k8s 的服务部署方法，其特征在于，所述获取所述第一目标 pod 对应的预设 ConfigMap 之后，还包括：

当所述 Proxy 接收到待传输数据时，通过所述服务对所述待传输数据进行处理，得到处理后待传输数据；

通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod，其中，所述第二目标 pod 为根据所述预设 ConfigMap 中的路由信息确定出的。

25 5、根据权利要求 4 所述的基于 k8s 的服务部署方法，其特征在于，所述通过所述 Proxy 将所述处理后待传输数据发送给第二目标 pod 之前，还包括：

基于 HTTP 协议建立所述第一目标 pod 中的 Proxy 与所述第二目标 pod 中的 Proxy 之间的通信连接。

6、根据权利要求4所述的基于k8s的服务部署方法，其特征在于，所述通过所述Proxy将所述处理后待传输数据发送给第二目标pod之前，还包括：

根据所述第一目标pod中的Proxy名称确定所述第一目标pod在所述
5 预设ConfigMap中的路由信息中所处的当前节点；

根据所述路由信息和所述第一目标pod所处的当前节点确定出所述第二目标pod。

7、根据权利要求6所述的数据传输方法，其特征在于，所述根据所述
10 路由信息和所述第一目标pod所处的当前节点确定出所述第二目标pod，
包括：

如果所述第一目标pod所处的当前节点为所述路由信息中的最后一个节点，则将所述路由信息中的第一个节点确定为所述第二目标pod。

8、一种基于k8s的服务部署装置，其特征在于，包括：

第一数据获取模块，用于获取服务数据，其中，所述服务数据中包括
15 一个服务对应的数据；

第一部署模块，用于根据所述服务数据在第一目标pod上部署服务；

第二部署模块，用于基于sidecar方式在所述第一目标pod上部署所述
服务对应的Proxy，以便将所述Proxy作为所述服务的网络代理；

第二数据获取模块，用于获取所述第一目标pod对应的预设
20 ConfigMap，以便所述Proxy根据所述预设ConfigMap进行不同pod之间的
通信，其中，所述预设ConfigMap中包括路由信息。

9、一种电子设备，其特征在于，包括：

存储器和处理器；

其中，所述存储器，用于存储计算机程序；

25 所述处理器，用于执行所述计算机程序，以实现权利要求1至7任一
项所述的基于k8s的服务部署方法。

10、一种计算机可读存储介质，其特征在于，用于保存计算机程序，
其中，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述
的基于k8s的服务部署方法。

— 1/5 —

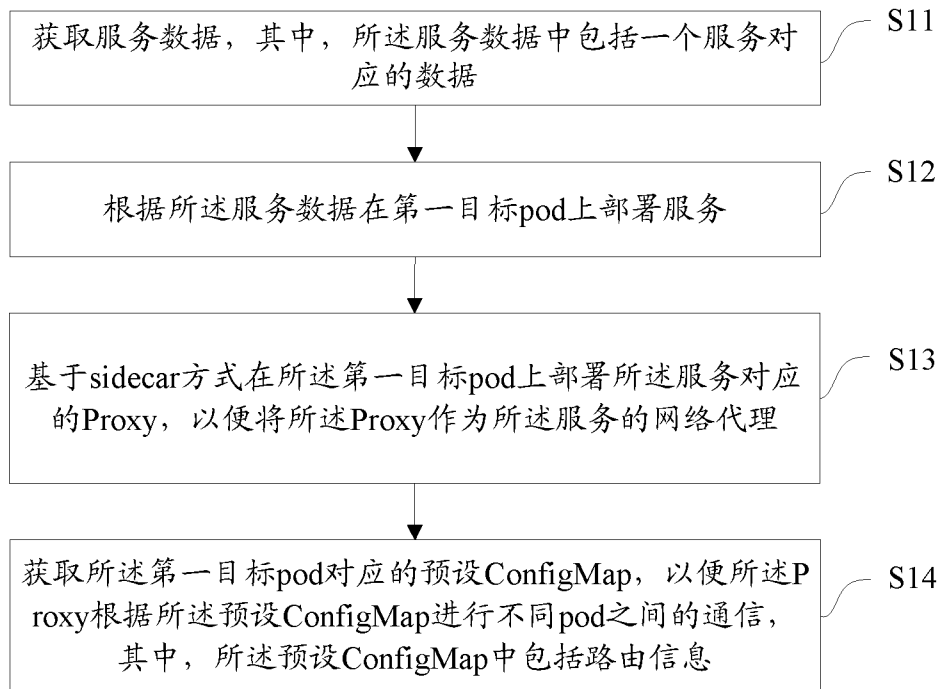


图 1

—2/5—

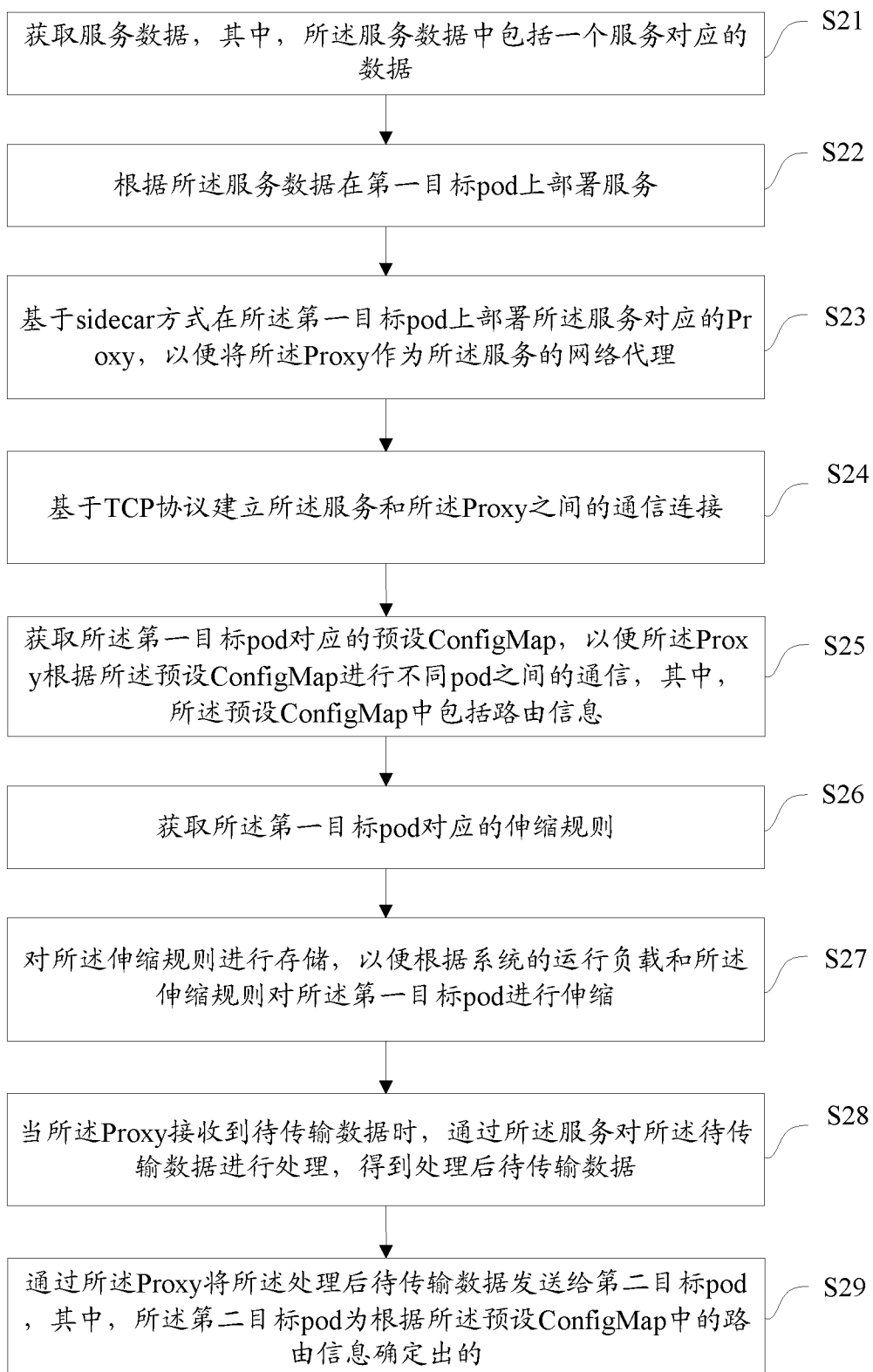


图 2

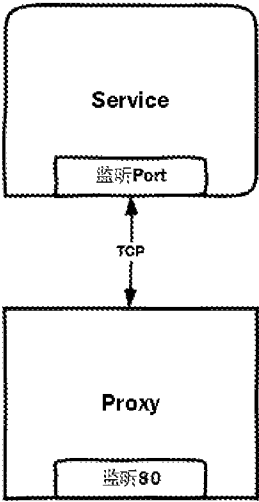


图 3

```
configmap.yaml

RouteMap:

  name: proxyA
  type: START
  addr: http://ip:port/methodA
  next: proxyB

  name: proxyB
  type: PROCESS
  addr: http://ip:port/methodB
  next: proxyC

  name: proxyC
  type: END
  addr: http://ip:port/methodC
  next: NULL
```

图 4

—4/5—

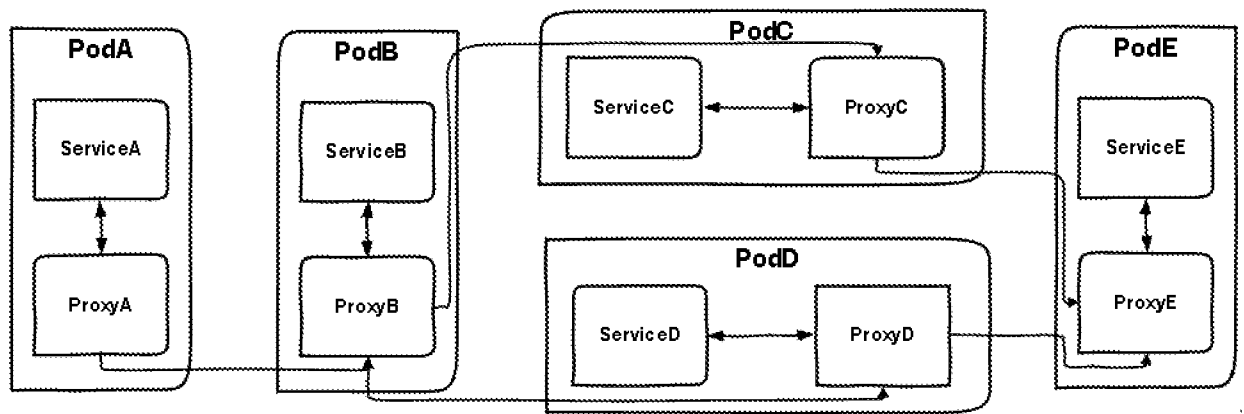


图 5

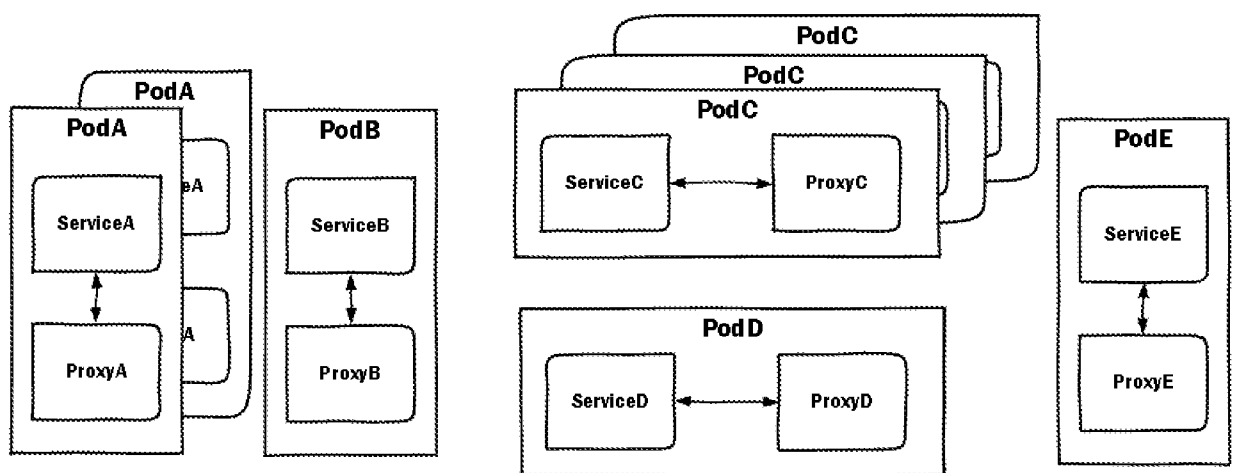


图 6

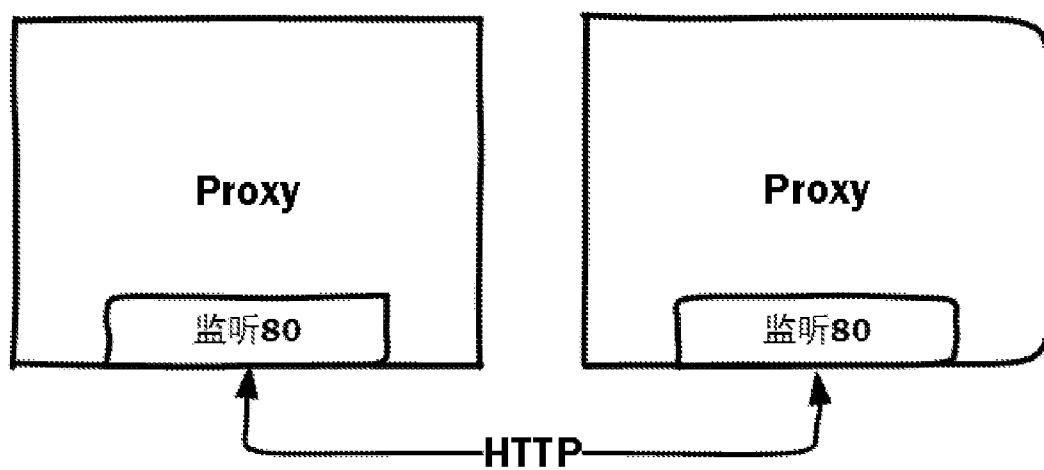


图 7

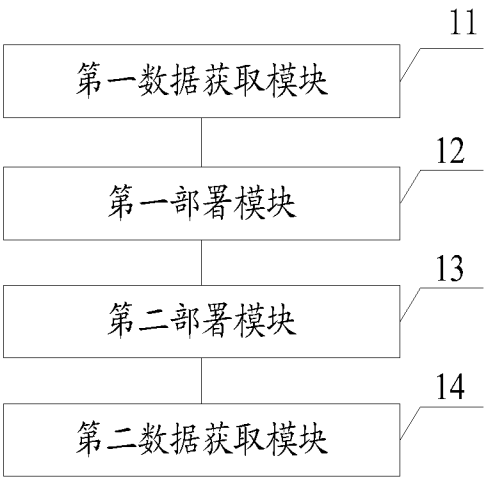


图 8

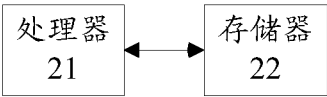


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/076995

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/24(2006.01)i

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)

H04L; G06F

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

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

CNPAT, WPI, EPODOC, CNKI, IEEE: 容器云, 路由, 代理, 边车, 配置, 部署单元, kubernetes, k8s, container, route, proxy, sidecar, config+, configmap, pod

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 111901157 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 06 November 2020 (2020-11-06) claims 1-10	1-10
X	CN 111212129 A (BEIJING INSPUR DATA TECHNOLOGY CO., LTD.) 29 May 2020 (2020-05-29) description, paragraphs [0062]-[0086]	1-10
A	CN 110716744 A (TECHNOLOGY AND ENGINEERING CENTER FOR SPACE UTILIZATION, CHINESE ACADEMY OF SCIENCES) 21 January 2020 (2020-01-21) entire document	1-10
A	CN 109981789 A (SHANDONG INSPUR CLOUD INFORMATION TECHNOLOGY CO., LTD.) 05 July 2019 (2019-07-05) entire document	1-10
A	US 10528367 B1 (INTUIT INC.) 07 January 2020 (2020-01-07) entire document	1-10
A	BALLA, David et al. "Adaptive scaling of Kubernetes pods" IEEE, 08 June 2020 (2020-06-08), entire document	1-10



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

"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

28 April 2021

Date of mailing of the international search report

21 May 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/076995

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	111901157	A	06 November 2020	None	
CN	111212129	A	29 May 2020	None	
CN	110716744	A	21 January 2020	None	
CN	109981789	A	05 July 2019	None	
US	10528367	B1	07 January 2020	None	

国际检索报告

国际申请号

PCT/CN2021/076995

A. 主题的分类 H04L 12/24 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																							
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) H04L; G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, WPI, EPDOC, CNKI, IEEE: 容器云, 路由, 代理, 边车, 配置, 部署单元, kubernetes, k8s, container, route, proxy, sidecar, config+, configmap, pod																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 111901157 A (苏州浪潮智能科技有限公司) 2020年 11月 6日 (2020 - 11 - 06) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>X</td> <td>CN 111212129 A (北京浪潮数据技术有限公司) 2020年 5月 29日 (2020 - 05 - 29) 说明书第[0062]-[0086]段</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 110716744 A (中国科学院空间应用工程与技术中心) 2020年 1月 21日 (2020 - 01 - 21) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 109981789 A (山东浪潮云信息技术有限公司) 2019年 7月 5日 (2019 - 07 - 05) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>US 10528367 B1 (INTUIT INC.) 2020年 1月 7日 (2020 - 01 - 07) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>BALLA, David等. "Adaptive scaling of Kubernetes pods" IEEE, 2020年 6月 8日 (2020 - 06 - 08), 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 111901157 A (苏州浪潮智能科技有限公司) 2020年 11月 6日 (2020 - 11 - 06) 权利要求1-10	1-10	X	CN 111212129 A (北京浪潮数据技术有限公司) 2020年 5月 29日 (2020 - 05 - 29) 说明书第[0062]-[0086]段	1-10	A	CN 110716744 A (中国科学院空间应用工程与技术中心) 2020年 1月 21日 (2020 - 01 - 21) 全文	1-10	A	CN 109981789 A (山东浪潮云信息技术有限公司) 2019年 7月 5日 (2019 - 07 - 05) 全文	1-10	A	US 10528367 B1 (INTUIT INC.) 2020年 1月 7日 (2020 - 01 - 07) 全文	1-10	A	BALLA, David等. "Adaptive scaling of Kubernetes pods" IEEE, 2020年 6月 8日 (2020 - 06 - 08), 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 111901157 A (苏州浪潮智能科技有限公司) 2020年 11月 6日 (2020 - 11 - 06) 权利要求1-10	1-10																					
X	CN 111212129 A (北京浪潮数据技术有限公司) 2020年 5月 29日 (2020 - 05 - 29) 说明书第[0062]-[0086]段	1-10																					
A	CN 110716744 A (中国科学院空间应用工程与技术中心) 2020年 1月 21日 (2020 - 01 - 21) 全文	1-10																					
A	CN 109981789 A (山东浪潮云信息技术有限公司) 2019年 7月 5日 (2019 - 07 - 05) 全文	1-10																					
A	US 10528367 B1 (INTUIT INC.) 2020年 1月 7日 (2020 - 01 - 07) 全文	1-10																					
A	BALLA, David等. "Adaptive scaling of Kubernetes pods" IEEE, 2020年 6月 8日 (2020 - 06 - 08), 全文	1-10																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 引用文件的具体类型: "A" 认为不特别相关的表示了现有技术一般状态的文件 "E" 在国际申请日的当天或之后公布的在先申请或专利 "L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) "O" 涉及口头公开、使用、展览或其他方式公开的文件 "P" 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> "T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 "X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 "Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 "&" 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: "A" 认为不特别相关的表示了现有技术一般状态的文件 "E" 在国际申请日的当天或之后公布的在先申请或专利 "L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) "O" 涉及口头公开、使用、展览或其他方式公开的文件 "P" 公布日先于国际申请日但迟于所要求的优先权日的文件	"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 "X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 "Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 "&" 同族专利的文件																			
* 引用文件的具体类型: "A" 认为不特别相关的表示了现有技术一般状态的文件 "E" 在国际申请日的当天或之后公布的在先申请或专利 "L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) "O" 涉及口头公开、使用、展览或其他方式公开的文件 "P" 公布日先于国际申请日但迟于所要求的优先权日的文件	"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 "X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 "Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 "&" 同族专利的文件																						
国际检索实际完成的日期 2021年 4月 28日		国际检索报告邮寄日期 2021年 5月 21日																					
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 寇利敏 电话号码 86-(10)-53961731																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/076995

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	111901157	A	2020年 11月 6日	无	
CN	111212129	A	2020年 5月 29日	无	
CN	110716744	A	2020年 1月 21日	无	
CN	109981789	A	2019年 7月 5日	无	
US	10528367	B1	2020年 1月 7日	无	