

(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
2017 年 3 月 16 日 (16.03.2017)



(10) 国际公布号
WO 2017/041650 A1

- (51) 国际专利分类号:
H04L 29/08 (2006.01)
- (21) 国际申请号: PCT/CN2016/097494
- (22) 国际申请日: 2016 年 8 月 31 日 (31.08.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510570429.9 2015 年 9 月 9 日 (09.09.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 成柱石 (CHENG, Zhushi); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。陶云峰 (TAO, Yun-feng); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。卢毅军 (LU, Yijun); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。
- (74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

[见续页]

(54) Title: METHOD AND DEVICE FOR EXTENDING DISTRIBUTED CONSISTENCY SERVICE

(54) 发明名称: 用于扩展分布式一致性服务的方法和设备

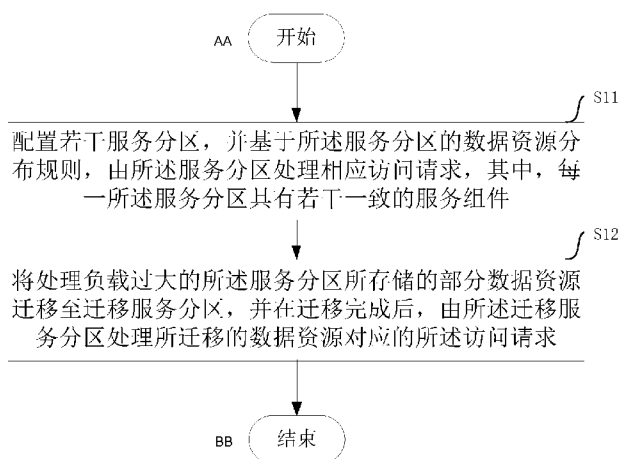


图 1

S11 Configuring several service partitions and on the basis of the data resource distribution rules of said service partitions, processing, by the service partitions, the corresponding access request, each of service partitions having several consistent service components

S12 Migrating the part of the data resources stored in said service partition with excessive processing load to the migration service partition and processing, by the migration service partition, the access request corresponding to the migrated data resources after the migration is completed

AA Start
BB End

(57) Abstract: The present invention provides a method and a device for extending distributed consistency service, in which a configuration device configures several service partitions having a consistency service and sets distribution rules for data resources served by each service partition; in this way, data resources are distributed to a plurality of service partitions, where each of service partitions processes a part of access requests and stores a part of the corresponding data resources; meanwhile, when the processing load of one or several service partitions is too heavy due to increases in access requests or data resources, a migration service partition is dynamically configured by means of a migration device, and then the part of the data resources stored in said service partition with excessive processing load are migrated to the migration service partition, and the migration service partition processes the access request corresponding to the migrated data resources after the migration is completed, thereby balancing the load pressure among the plurality of service partitions and ensuring service stability.

(57) 摘要: 本申请提供了用于扩展分布式一致性服务的方法和设备, 通过配置装置配置若干具有一致性服务的分区, 并设定每一分区所服务的分区的数据资源的分布规则, 从而将数据资源分散到多个分区中, 每个分区处理一部分访问请求并存储部分相应数据资源; 同时, 当访问请求增多或数据资源增大时, 造成某个或某几个分区的处理负载过大时, 则通过迁移装置动态配置迁移服务分区, 然后将处理负载过大的所述分区所存储的部分数据资源迁移至迁移服务分区, 并在迁移完成后, 由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求, 从而平衡多个分区之间的负载压力, 保证服务稳定性。

据资源增大时, 造成某个或某几个服务分区的处理负载过大时, 则通过迁移装置动态配置迁移服务分区, 然后将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区, 并在迁移完成后, 由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求, 从而平衡多个服务分区之间的负载压力, 保证服务稳定性。



WO 2017/041650 A1



(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))。

用于扩展分布式一致性服务的方法和设备

本申请要求 2015 年 09 月 09 日递交的申请号为 201510570429.9 发明名称为“用于扩展分布式一致性服务的方法和设备”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

本申请涉及计算机领域，尤其涉及一种用于扩展分布式一致性服务的方法和设备。

背景技术

10 在分布式系统中，通常由多个服务组件共同接受访问请求并提供服务，以保持稳定性，同时使用分布式一致性服务（例如分布式锁）来协调多个服务组件之间的关系和操作。当前主流分布式锁都会基于 quorum（多数派机制，例如 chubby、zookeeper）来保证一致性和可用性。现有基于 quorum 实现的分布式锁，一般会面临下面几个严重的问题：

每次更新操作，需要经过半数以上服务组件同意，才可以认为一次更新操作成功。
15 当更新操作非常频繁时，没有固有的办法能够简单有效地提高系统的服务能力。典型的，简单的添加服务组件的方法，一方面不能提高服务的处理能力，同时还会影响更新操作的响应时间。此外，分布式锁对自身服务的失效时间（Failover）有着异常苛刻的要求，当服务的数据规模逐渐增大时，访问请求越来越多，存储的数据资源越来越大，失效时间会随之同步增长，服务的可用性会受到非常大的挑战。

20

发明内容

本申请要解决的技术问题是：提供一种基于服务分区的扩展分布式一致性服务的方法和设备。

根据本申请一方面提供的一种用于扩展分布式一致性服务的方法，其中，所述方法
25 包括：

配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

30 进一步地，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务

分区包括：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；

将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。

进一步地，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务

5 分区还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；

在即将完成迁移时，将所述迁移前数据资源设置为只读状态；

在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态。

10 进一步地，在迁移完成后由所述迁移服务分区处理与所迁移的数据资源相应的所述访问请求包括：

更新所述数据资源分布规则；

向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；

15 在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。

进一步地，所述标识文件包括分布式锁文件。

进一步地，所述服务分区的服务组件采用多数派机制维持一致性。

进一步地，所述迁移服务分区为已有空闲的服务分区或新配置的服务分区。

20 根据本申请一方面提供的一种用于扩展分布式一致性服务的设备，其中，所述设备包括：

配置装置，用于配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

25 迁移装置，用于将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

进一步地，所述迁移装置用于：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；

将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。

30 进一步地，所述迁移装置还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；

在即将完成迁移时，将所述迁移前数据资源设置为只读状态；

在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态。

- 5 进一步地，所述迁移装置用于：在迁移完成后，
更新所述数据资源分布规则；

向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；

- 10 在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。

进一步地，所述标识文件包括分布式锁文件。

进一步地，所述迁移装置用于：

重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态后超出超时时间时，删除所述迁移前数据资源。

- 15 进一步地，所述服务分区的服务组件采用多数派机制维持一致性。

进一步地，所述迁移服务分区为已有空闲的服务分区或新配置的服务分区。

- 与现有技术相比，在本申请一方面提供的所述用于扩展分布式一致性服务的方法及另一方面提供的设备，通过部署多组包括分布式锁服务组件的服务分区、每组服务分区存储部分数据资源，并提供相应访问请求，并在运行时动态调整每个分区服务的数据资源来平衡各组服务分区之间的处理负载压力，以此提高了分布式一致性服务的扩展能力。
- 20

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

- 25 图 1 示出根据本申请一个方面的提供的一种用于扩展分布式一致性服务的方法流程图；

图 2~图 4 示出本申请一个优选实施例提供的一种用于扩展分布式一致性服务的迁移过程示意图；

- 30 图 5 示出根据本申请一个方面的提供的一种用于扩展分布式一致性服务的设备示意图。

附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

下面结合附图对本申请作进一步详细描述。

5 在分布式系统中，分布式一致性服务一般基于 C/S 架构（客户/服务架构）来实现，服务一般由多个服务组件形成的多个节点提供服务，以保证一致性和可用性；客户端提供接口给用户或应用方用来访问分布式一致性服务。分布式一致性服务包括分布式锁，分布式锁是控制分布式系统之间同步访问共享资源的一种方式。如果不同的系统或是同一个系统的不同主机之间共享了一个或一组服务组件的数据资源，那么请求访问这些数
10 据资源的时候，往往需要互斥来防止彼此干扰来保证一致性，在这种情况下，便需要使用到分布式锁。

 基于以上，本申请一方面提供了一种基于服务分区的分布式一致性服务的方法，通过配置若干具有一致性服务的服务分区，并设定每一服务分区所服务的数据资源的分布规则，从而将数据资源分散到多个服务分区中，每个服务分区处理一部分访问请求并存储部分相应数据资源；同时，当访问请求增多或数据资源增大时，造成某个或某几个服务分区的处理负载过大时，则动态配置迁移服务分区，迁移服务分区可以是已有空闲的服务分区或新配置的服务分区，然后将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求，从而平衡多个服务分区之间的负载压力，保证服务稳定性。

20 图 1 示出根据本申请一个方面的提供的一种用于扩展分布式一致性服务的方法流程图，其中，所述方法包括：

 步骤 S11：配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

 步骤 S12：将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。
25

 在所述步骤 S11 中，所述服务分区的服务组件采用多数派机制维持一致性，相同所述服务分区中的服务组件存储相同的数据资源。在配置完成后，当访问请求到达分布式一致性服务前端时，会基于访问请求的内容，并根据数据资源分布规则，将访问请求路由到相应能够处理该访问请求的服务分区中，由该服务分区完成对用户的访问申请的处
30

理、释放分布式锁，设置分布式锁的拥有者信息等操作。

在所述步骤 S12 中，在分布式一致性服务工作一段时间后，某个或某些服务分区上会出现访问请求过多或存储数据资源过大，导致处理负载过大，此时可以根据系统的运行情况，动态将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

进一步地，在所述步骤 S12 中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区包括：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；具体地，选定方式并不被限制，可根据数据资源的数据量及所对应的访问请求对服务分区进行具体划分；将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。以复制的方式能够保证在迁移过程中，不影响迁移前已拥有标识文件（分布式锁文件）的用户继续提供数据服务。

进一步地，在所述步骤 S12 中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；设置为无效状态用以防止系统认为迁移后数据资源超时时间内未获得访问请求而删除所述迁移后数据资源。

在即将完成迁移时，将所述迁移前数据资源设置为只读状态；此时，用户已经持有的标识文件并不受影响，即可以继续利用所述标识文件读取迁移前数据资源，

在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态，为数据资源 resource-x 开始提供服务做准备。

进一步地，在所述步骤 S12 中，在迁移完成后由所述迁移服务分区处理与所迁移的数据资源相应的所述访问请求包括：

更新所述数据资源分布规则；

向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；具体地，若收到来自拥有迁移前数据资源的标识文件的用户的拥有者变更操作请求后，会把对应迁移后数据资源中的标识文件设置为有效、且将拥有者信息更改为该原拥有者用户，若作为拥有迁移前数据资源的标识文件的用户在收到通知后未向服务分区发起拥有者请求操作，在等待超出预设时间后，则删除相应标识文件，可以根据其他用户的访问请求重新释放分布式锁。

在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。其中，优选地，所述标识文件包括分布式锁文件。

此外，当重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态后
5 超出超时时间时，删除所述迁移前数据资源。，最终完成数据迁移。

图 2~图 4 示出本申请一个优选实施例提供的一种扩展分布式一致性服务的迁移过程示意图，在具体的场景中，以服务分区 quorum-n 为例，来说明用于扩展分布式一致性服务的方法中动态迁移数据的过程：

如图 2 所示，初始配置后，例如，在提供服务的服务分区包括服务分区 quorum-1 至
10 服务分区 quorum-n，其中，服务分区 quorum-n 上有数据资源 resource-n 和数据资源 resource-x，都由服务分区 quorum-n 服务，在 quorum-n 服务为用户分配关于数据资源 resource-n 的锁文件 lock://resource-n/foo，为用户分配关于数据资源 resource-x 的锁文件 lock://resource-x/foo，作为拥有者的用户可以持所述锁文件 lock://resource-n/foo 获取数据资源 resource-n 的相关服务，持所述锁文件 lock://resource-x/foo 获取数据资源 resource-x
15 的相关服务。

如图 3 所示，当服务分区 quorum-n 的处理负载压力过大，决定把数据资源 resource-x 迁移到较空闲的服务分区 quorum-x 上。

首先，启动一个迁移操作（migration 操作），实时地将服务分区 quorum-n 上和
数据资源 resource-x 相关的数据动态以复制的方式迁移到迁移服务分区 quorum-x 上。
20 Migration 操作利用分布式锁原生的同步协议，可以实时的同步所有数据及更新操作到新的分区，整个过程对原有服务不会产生任何影响。此时，迁移到迁移服务分区 quorum-x 上的迁移后数据资源 resource-x' 设置为无效状态（invalid），设置为无效状态用以防止系统认为迁移后数据资源 resource-x' 超时时间内未获得访问请求而删除所述迁移后数据资源 resource-x' 。

25 当迁移即将完成时，服务分区 quorum-n 的数据资源 resource-x 和服务分区 quorum-x 上数据资源 resource-x' 的相关数据的状态基本一致时，例如需要迁移的数据资源一共有 1024M，还剩 8M 需要迁移的数据需要传送，剩余需要迁移的数据资源所需传递的时间远小于一次访问请求的处理时间时，在服务分区 quorum-n 上对数据资源 resource-x 设置只读状态（read-only）。此时，用户已经持有的锁文件 lock://resource-x/foo 并不受影响，
30 即可以继续利用所述锁文件读取数据资源 resource-x。

当迁移完成时，待服务分区 quorum-n 的数据资源 resource-x 和服务分区 quorum-x 上数据资源 resource-x' 的相关数据的完全一致后，则在服务分区 quorum-n 上禁止所有对数据资源 resource-x 的操作，同时在服务分区 quorum-x 上设置数据资源 resource-x' 为只读状态（read-only），即为数据资源 resource-x' 开始提供服务做准备。

- 5 更新所有服务分区路由规则：数据资源 resource-x' 由服务分区 quorum-x 服务，数据资源 resource-x' 与数据资源 resource-x 完全一致。

通过系统前端通知所有用户作拥有者变更操作（chown 操作）。作为原拥有者的用户在收到通知后会向服务分区 quorum-x 发起一个 chown 操作。则服务分区 quorum-x 收到来自原拥有者的用户的 chown 操作后，会把对应的锁文件（lock://resource-x/foo）设置
10 为有效、且拥有者信息更改为该原拥有者用户，若作为原拥有者的用户在收到通知后未向服务分区 quorum-x 发起 chown 操作，在等待超出预设时间后，则删除服务分区 quorum-x 上未被 chown 操作的锁文件 lock://resource-x/foo，可以根据其他用户的访问请求重新释放分布式锁，同时，在超出预设时间后，取消数据资源 resource-x 和数据资源 resource-x' 的状态限制，即均为可读可写状态，在超出超时时间内数据资源 resource-x
15 未收到访问请求的心跳报文，则系统会清除数据资源 resource-x，最终完成数据迁移，如图 4 所示。

当然，本领域技术人员应能理解上述分布式一致性服务的动态迁移的具体场景及其中锁文件、服务分区、数据资源、chown 操作，migration 操作的描述仅为举例，其他现有的或今后可能出现的其他描述如可适用于本申请，也应包含在本申请保护范围以内，
20 并在此以引用方式包含于此。

图 5 示出根据本申请一个方面的提供的一种用于扩展分布式一致性服务的设备示意图，其中，所述设备 1 包括：

配置装置 11，配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

- 25 迁移装置 12，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

在此，所述设备 1 包括但不限于用户设备、或用户设备与网络设备通过网络相集成所构成的设备。所述用户设备其包括但不限于任何一种可与用户通过触摸板进行人机交互的移动电子产品，例如智能手机、PDA 等，所述移动电子产品可以采用任意操作系统，
30

如 android 操作系统、iOS 操作系统等。其中，所述网络设备包括一种能够按照事先设定或存储的指令，自动进行数值计算和信息处理的电子设备，其硬件包括但不限于微处理器、专用集成电路(ASIC)、可编程门阵列(FPGA)、数字处理器(DSP)、嵌入式设备等。

所述网络包括但不限于互联网、广域网、城域网、局域网、VPN 网络、无线自组织网络(Ad Hoc 网络)等。优选地，设备 1 还可以是运行于所述用户设备、或用户设备与网络设备、触摸终端或网络设备与触摸终端通过网络相集成所构成的设备上的脚本程序。当然，本领域技术人员应能理解上述设备 1 仅为举例，其他现有的或今后可能出现的设备 1 如可适用于本申请，也应包含在本申请保护范围以内，并在此以引用方式包含于此。

上述各装置之间是持续不断工作的，在此，本领域技术人员应理解“持续”是指上述各装置分别实时地或者按照设定的或实时调整的工作模式要求，直至所述设备 1 停止工作。

所述配置装置 11 对所述服务分区的服务组件采用多数派机制维持一致性，相同所述服务分区中的服务组件存储相同的数据资源。在配置完成后，当访问请求到达分布式一致性服务的前端时，会基于访问请求的内容，并根据数据资源分布规则，将访问请求路由到相应能够处理该访问请求的服务分区中，由该服务分区完成对用户的访问申请的处理、释放分布式锁，设置分布式锁的拥有者信息等操作。

在所述迁移装置 12 中，在分布式一致性服务在工作一段时间中，某个或某些服务分区上会出现访问请求过多或存储数据资源过大，导致处理负载过大，此时可以根据系统的运行情况，动态将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

进一步地，在所述迁移装置 12 中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区包括：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；具体地，选定方式并不被限制，可根据数据资源的数据量及所对应的访问请求对服务分区进行具体划分；将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。以复制的方式能够保证在迁移过程中，不影响迁移前已拥有标识文件（分布式锁文件）的用户继续提供数据服务。

进一步地，在所述迁移装置 12 中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；设置为无效状态用以防止系统认为迁移后数据资源超时时间内未获得访问请求而删除所述迁移后数据资源。

在即将完成迁移时，将所述迁移前数据资源设置为只读状态；此时，用户已经持有的标识文件并不受影响，即可以继续利用所述标识文件读取迁移前数据资源，

5 在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态，为数据资源 resource-x 开始提供服务做准备。

进一步地，在所述迁移装置 12 中，在迁移完成后由所述迁移服务分区处理与所迁移的数据资源相应的所述访问请求包括：

更新所述数据资源分布规则；

10 向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；具体地，若收到来自拥有迁移前数据资源的标识文件的用户的拥有者变更操作请求后，会把对应迁移后数据资源中的标识文件设置为有效、且将拥有者信息更改为该原拥有者用户，若作为拥有迁移前数据资源的标识文件的用户在收到通知后未向服务分区发起拥有者请求操作，在等待超出预设时间后，则删除相应标识文件，可以根据其他用户的访问请求重新释放分布式锁。

在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。其中，优选地，所述标识文件包括分布式锁文件。

20 图 2~图 4 示出本申请一个优选实施例提供的一种扩展分布式一致性服务的迁移过程示意图，在具体的场景中，以服务分区 quorum-n 为例，来说明用于扩展分布式一致性服务的设备 1 进行动态迁移数据的过程：

如图 2 所示，初始配置后，例如，在服务分区 quorum-n 上有数据资源 resource-n 和数据资源 resource-x，都由服务分区 quorum-n 服务，在 quorum-n 服务为用户分配关于数据资源 resource-n 的锁文件 lock://resource-n/foo，为用户分配关于数据资源 resource-x 的锁文件 lock://resource-x/foo，作为拥有者的用户可以持所述锁文件 lock://resource-n/foo 获取数据资源 resource-n 的相关服务，持所述锁文件 lock://resource-x/foo 获取数据资源 resource-x 的相关服务。

30 如图 3 所示，当服务分区 quorum-n 的处理负载压力过大，决定把数据资源 resource-x 迁移到较空闲的服务分区 quorum-x 上。

首先，启动一个迁移操作（migration 操作），实时地将服务分区 quorum-n 上和数据资源 resource-x 相关的数据动态以复制的方式迁移到迁移服务分区 quorum-x 上。Migration 操作利用分布式锁原生的同步协议，可以实时的同步所有数据及更新操作到新的分区，整个过程对原有服务不会产生任何影响。此时，迁移到迁移服务分区 quorum-x 上的迁移后数据资源 resource-x' 设置为无效状态（invalid），设置为无效状态用以防止系统认为迁移后数据资源 resource-x' 超时时间内未获得访问请求而删除所述迁移后数据资源 resource-x'。

当迁移即将完成时，服务分区 quorum-n 的数据资源 resource-x 和服务分区 quorum-x 上数据资源 resource-x' 的相关数据的状态基本一致时，例如需要迁移的数据资源一共有 1024M，还剩 8M 需要迁移的数据需要传送，剩余需要迁移的数据资源所需传递的时间远小于一次访问请求的处理时间时，在服务分区 quorum-n 上对数据资源 resource-x 设置只读状态（read-only）。此时，用户已经持有的锁文件 lock://resource-x/foo 并不受影响，即可以继续利用所述锁文件读取数据资源 resource-x。

当迁移完成时，待服务分区 quorum-n 的数据资源 resource-x 和服务分区 quorum-x 上数据资源 resource-x' 的相关数据的完全一致后，则在服务分区 quorum-n 上禁止所有对数据资源 resource-x 的操作，同时在服务分区 quorum-x 上设置数据资源 resource-x' 为只读状态（read-only），即为数据资源 resource-x' 开始提供服务做准备。

更新所有服务分区路由规则：数据资源 resource-x' 由服务分区 quorum-x 服务，数据资源 resource-x' 与数据资源 resource-x 完全一致。

通过系统前端通知所有用户作拥有者变更操作（chown 操作）。作为原拥有者的用户在收到通知后会向服务分区 quorum-x 发起一个 chown 操作。则服务分区 quorum-x 收到来自原拥有者的用户的 chown 操作后，会把对应的锁文件（lock://resource-x/foo）设置为有效、且拥有者信息更改为该原拥有者用户，若作为原拥有者的用户在收到通知后未向服务分区 quorum-x 发起 chown 操作，在等待超出预设时间后，则删除服务分区 quorum-x 上未被 chown 操作的锁文件 lock://resource-x/foo，可以根据其他用户的访问请求重新释放分布式锁，同时，在超出预设时间后，取消数据资源 resource-x 和数据资源 resource-x' 的状态限制，即均为可读可写状态，在超出超时时间内数据资源 resource-x 未收到访问请求的心跳报文，则系统会清除数据资源 resource-x，最终完成数据迁移，如图 4 所示。

当然，本领域技术人员应能理解上述扩展分布式一致性服务的动态迁移的具体场景

及其中锁文件、服务分区、数据资源、chown 操作，migration 操作的描述仅为举例，其他现有的或今后可能出现的其他描述如可适用于本申请，也应包含在本申请保护范围以内，并在此以引用方式包含于此。

与现有技术相比，根据本申请的实施例提供的基于服务分区的扩展分布式一致性服务的设备，通过配置装置配置若干具有一致性服务的服务分区，并设定每一服务分区所服务的数据资源的分布规则，从而将数据资源分散到多个服务分区中，每个服务分区处理一部分访问请求并存储部分相应数据资源；同时，当访问请求增多或数据资源增大时，造成某个或某几个服务分区的处理负载过大时，则通过迁移装置动态配置迁移服务分区，然后将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求，从而平衡多个服务分区之间的负载压力，保证服务稳定性。

需要注意的是，本申请可在软件和/或软件与硬件的组合物中被实施，例如，可采用专用集成电路(ASIC)、通用目的计算机或任何其他类似硬件设备来实现。在一个实施例中，本申请的软件程序可以通过处理器执行以实现上文所述步骤或功能。同样地，本申请的软件程序(包括相关的数据结构)可以被存储到计算机可读记录介质中，例如，RAM 存储器，磁或光驱动器或软磁盘及类似设备。另外，本申请的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

另外，本申请的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本申请的方法和/或技术方案。而调用本申请的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据所述程序指令运行的计算机设备的工作存储器中。在此，根据本申请的一个实施例包括一个装置，该装置包括用于存储计算机程序指令的存储器和用于执行程序指令的处理器，其中，当该计算机程序指令被该处理器执行时，触发该装置运行基于前述根据本申请的多个实施例的方法和/或技术方案。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权

利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。装置权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神
5 和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

权利要求书

1. 一种用于扩展分布式一致性服务的方法，其中，所述方法包括：

配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

5 将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

2. 根据权利要求 1 所述的方法，其中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区包括：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；

10 将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。

3. 根据权利要求 2 所述的方法，其中，将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；

15 在即将完成迁移时，将所述迁移前数据资源设置为只读状态；

在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态。

4. 根据权利要求 3 所述的方法，其中，在迁移完成后由所述迁移服务分区处理与所迁移的数据资源相应的所述访问请求包括：

20 更新所述数据资源分布规则；

向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；

在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。

25 5. 根据权利要求 4 所述的方法，其中，所述标识文件包括分布式锁文件。

6. 根据权利要求 1 至 5 中任一项所述的方法，其中，所述服务分区的服务组件采用多数派机制维持一致性。

7. 根据权利要求 1 至 6 中任一项所述的方法，其中，所述迁移服务分区为已有空闲的服务分区或新配置的服务分区。

30 8. 一种用于扩展分布式一致性服务的设备，其中，所述设备包括：

配置装置，用于配置若干服务分区，并基于所述服务分区的数据资源分布规则，由所述服务分区处理相应访问请求，其中，每一所述服务分区具有若干一致的服务组件；

迁移装置，用于将处理负载过大的所述服务分区所存储的部分数据资源迁移至迁移服务分区，并在迁移完成后，由所述迁移服务分区处理所迁移的数据资源对应的所述访问请求。

9. 根据权利要求 8 所述的设备，其中，所述迁移装置用于：

将处理负载过大的所述服务分区所存储的部分数据资源选定为迁移前数据资源；

将所述迁移前数据资源以复制方式迁移到所述迁移服务分区，形成迁移后数据资源。

10. 根据权利要求 9 所述的设备，其中，所述迁移装置还包括：

在迁移过程中，将所述迁移后数据资源设置为无效状态；

在即将完成迁移时，将所述迁移前数据资源设置为只读状态；

在完成迁移后，将所述迁移前数据资源设置为不可读不可写状态，并将所述迁移后数据设置为只读状态。

11. 根据权利要求 10 所述的设备，其中，所述迁移装置用于：在迁移完成后，更新所述数据资源分布规则；

向所有用户发起服务分区变更通知，并基于所述用户的拥有者变更操作请求，将迁移后数据资源中的标识文件设置为有效，且拥有者信息设置为相应的请求用户；

在完成迁移超出预设时间后，重置所述迁移前数据资源和所述迁移后数据资源的状态为可读可写状态，并删除迁移数据中未获得更改用户者操作请求的相应标识文件。

12. 根据权利要求 11 所述的设备，其中，所述标识文件包括分布式锁文件。

13. 根据权利要求 8 至 12 中任一项所述的设备，其中，所述服务分区的服务组件采用多数派机制维持一致性。

14. 根据权利要求 8 至 13 中任一项所述的设备，其中，所述迁移服务分区为已有空闲的服务分区或新配置的服务分区。

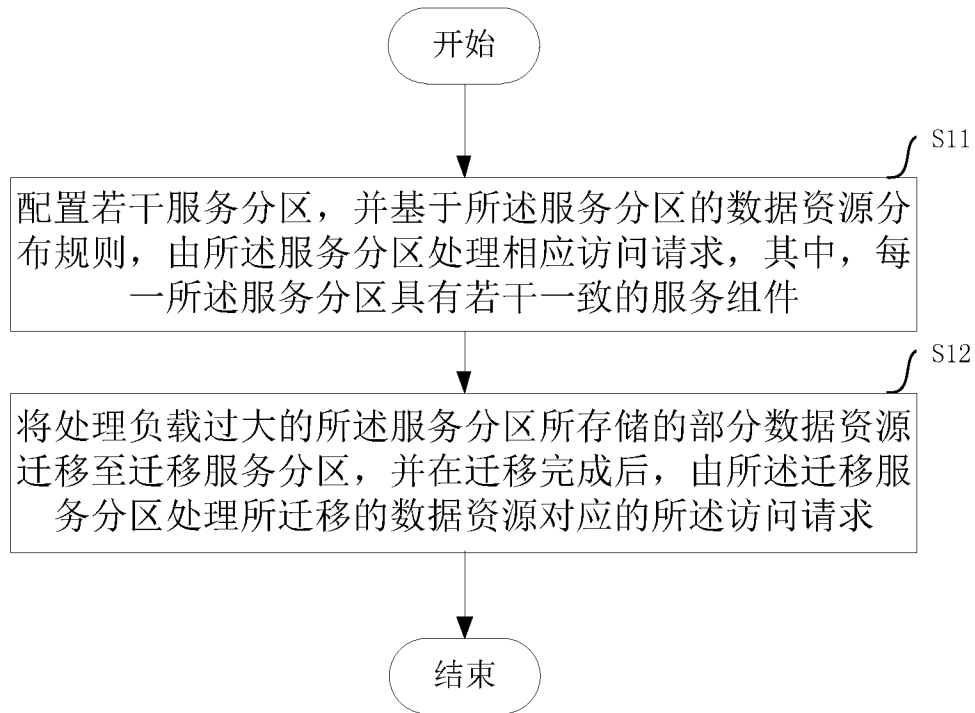


图 1

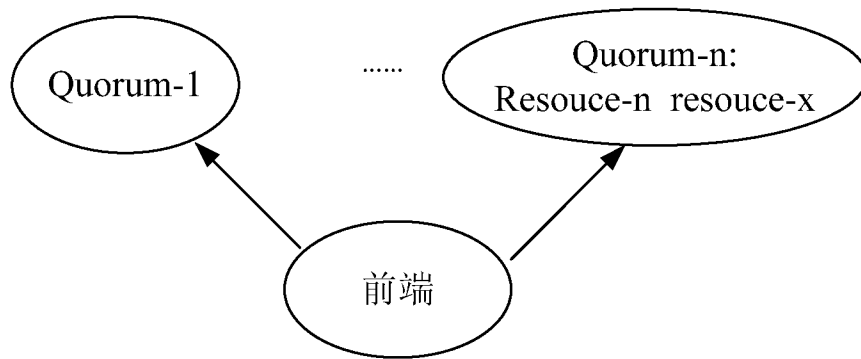


图 2

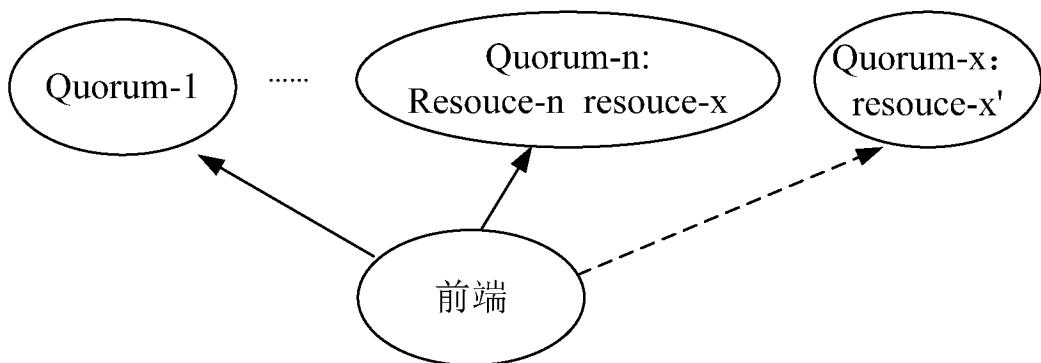


图 3

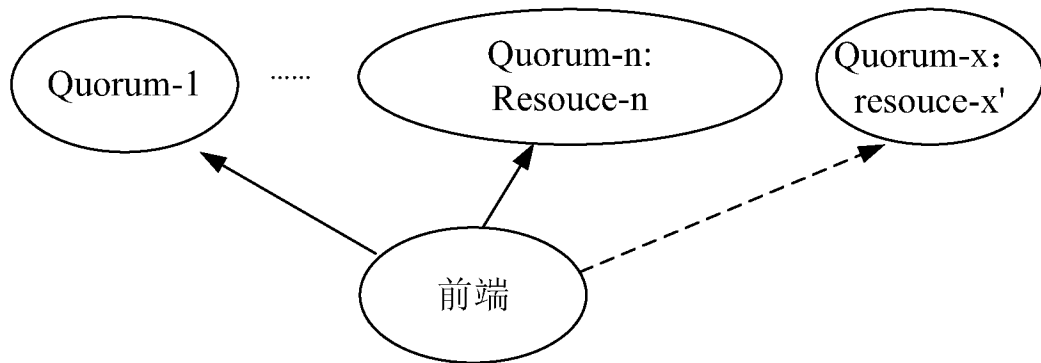


图 4

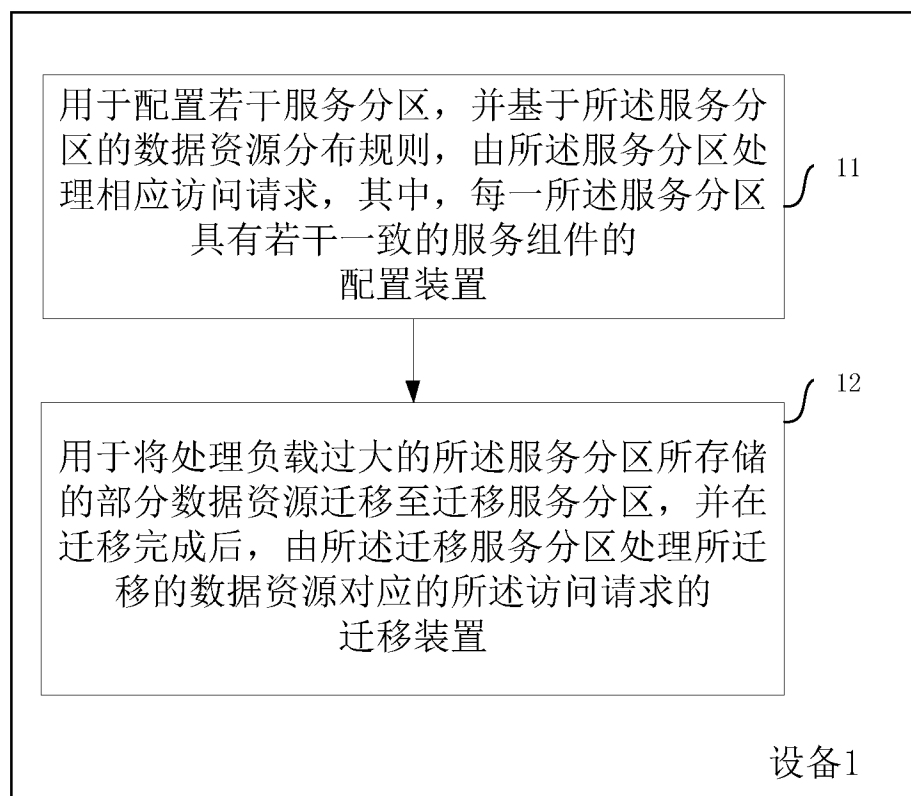


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/097494**A. CLASSIFICATION OF SUBJECT MATTER**

H04L 29/08 (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, CNKI, WPI, EPODOC: slice, service, data, store, storage, memory, visit, partition, block, load, transfer, distributed

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103327094 A (CHENGDU ONGOING INFORMATION TECHNOLOGY CO., LTD.), 25 September 2013 (25.09.2013), abstract, description, paragraphs 0005-0011 and 0014-0017, and figure 2	1-14
A	CN 101187949 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.), 28 May 2008 (28.05.2008), the whole document	1-14
A	CN 103036994 A (DAWNING INFORMATION INDUSTRY (BEIJING) CO., LTD.), 10 April 2013 (10.04.2013), the whole document	1-14
A	CN 101710339 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 19 May 2010 (19.05.2010), the whole document	1-14
A	CN 101909068 A (HUAWEI TECHNOLOGIES CO., LTD.), 08 December 2010 (08.12.2010), the whole document	1-14

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

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 03 November 2016 (03.11.2016)	Date of mailing of the international search report 25 November 2016 (25.11.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer JIN, Jing Telephone No.: (86-10) 62413414

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/CN2016/097494

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103327094 A	25 September 2013	None	
CN 101187949 A	28 May 2008	None	
CN 103036994 A	10 April 2013	None	
CN 101710339 A	19 May 2010	None	
CN 101909068 A	08 December 2010	None	

A. 主题的分类 H04L 29/08 (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, CNKI, WPI, EPDOC: 服务, 数据, 存储, 访问, 分区, 分片, 分块, 负载, 负荷, 迁移, 分布式, service, data, store, storage, memory, visit, partition, block, load, transfer, distributed		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 103327094 A (成都市欧冠信息技术有限责任公司) 2013年 9月 25日 (2013 - 09 - 25) 摘要、说明书第0005-0011段, 第0014-0017段、附图2	1-14
A	CN 101187949 A (腾讯科技深圳有限公司) 2008年 5月 28日 (2008 - 05 - 28) 全文	1-14
A	CN 103036994 A (曙光信息产业北京有限公司) 2013年 4月 10日 (2013 - 04 - 10) 全文	1-14
A	CN 101710339 A (中国科学院计算技术研究所) 2010年 5月 19日 (2010 - 05 - 19) 全文	1-14
A	CN 101909068 A (华为技术有限公司) 2010年 12月 8日 (2010 - 12 - 08) 全文	1-14
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 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 2016年 11月 3日		国际检索报告邮寄日期 2016年 11月 25日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 靳晶 电话号码 (86-10) 62413414

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/097494

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	103327094	A	2013年 9月 25日	无	
CN	101187949	A	2008年 5月 28日	无	
CN	103036994	A	2013年 4月 10日	无	
CN	101710339	A	2010年 5月 19日	无	
CN	101909068	A	2010年 12月 8日	无	