

(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
2025 年 2 月 20 日 (20.02.2025)



(10) 国际公布号
WO 2025/035366 A1

- (51) 国际专利分类号:
H04L 67/10 (2022.01)
- (21) 国际申请号: PCT/CN2023/112968
- (22) 国际申请日: 2023 年 8 月 14 日 (14.08.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 华为技术有限公司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
- (72) 发明人: 王君 (**WANG, Jun**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 彭程晖 (**PENG, Chenghui**); 中国广东省深

圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 王飞 (**WANG, Fei**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘哲 (**LIU, Zhe**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 黄欢欢 (**HUANG, Huanhuan**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

- (74) 代理人: 北京同达信恒知识产权代理有限公司 (**TDIP & PARTNERS**); 中国北京市西城区裕民路18号北环中心A座2002, Beijing 100029 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI,

(54) **Title:** RESOURCE SCHEDULING METHOD AND APPARATUS

(54) 发明名称: 一种资源调度方法和装置

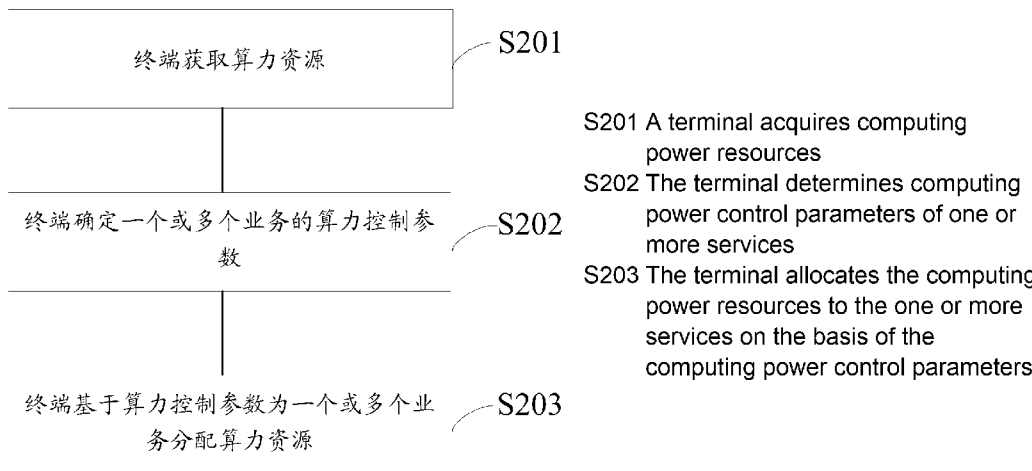


图 2

(57) **Abstract:** The present application relates to the technical field of wireless communications. Provided are a resource scheduling method and apparatus, which are used for realizing utilization of computing power resources by a terminal device. The method comprises: a terminal device acquiring computing power resources; and the terminal device determining computing power control parameters of one or more services, and allocating the computing power resources to the one or more services on the basis of the computing power control parameters. On the basis of the solution, a terminal device can allocate, on the basis of computing power control parameters, computing power resources to services having computing requirements, thus realizing the sharing and utilization of computing power of the terminal device.

(57) **摘要:** 本申请提供一种资源调度方法和装置, 用来实现终端设备对算力资源的利用, 涉及无线通信技术领域。该方法中, 终端设备获取算力资源。终端设备确定一个或多个业务的算力控制参数, 并基于算力控制参数为一个或多个业务分配算力资源。基于该方案, 终端设备可以基于算力控制参数为有计算需求的业务分配算力资源, 实现终端设备的算力的共享及利用。

GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,
IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ,
LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,
MU, 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) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, 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, ME, 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))。

一种资源调度方法和装置

技术领域

本申请涉及无线通信技术领域，尤其涉及一种资源调度方法和装置。

背景技术

伴随着人工智能（artificial intelligence，AI）超大模型，以及扩展现实（extended reality，XR）类等大计算量业务的出现，无论是网络设备或是终端设备，单点计算瓶颈都会逐步显现。因此在无线网络中支持算力共享是未来趋势。随着泛终端（如对于支持自动驾驶的以及具有不同等级的算力能力的车终端）的引入，终端设备的能力不断增强，终端可共享和利用其算力给网络从而辅助网络计算。目前尚无如何共享终端设备算力的技术方案。并且，相关技术中也没有终端在获取到算力资源后如何为任务分配算力的技术方案。

发明内容

本申请提供一种资源调度方法和装置，用来实现终端设备对算力共享及利用。

第一方面，提供一种资源调度方法。该方法可以由终端设备执行，或者芯片/芯片系统执行。该方法中，终端设备获取算力资源。终端设备确定一个或多个业务的算力控制参数，并基于算力控制参数为一个或多个业务分配算力资源。基于该方案，终端设备可以基于算力控制参数为有计算需求的业务分配算力资源，实现终端设备的算力的共享及利用。

在一种可能的实现方式中，终端设备接收来自网络设备的配置信息，该配置信息包括用于确定算力控制参数的参数。基于该方案，终端设备可以根据网络设备的配置确定算力控制参数。

在一种可能的实现方式中，配置信息包括保证时间窗和第一速率，算力控制参数是第一速率和保证时间窗的乘积，第一速率是单位时间内分配的算力大小，保证时间窗指用于确定算力控制参数的时长。

基于该方案，算力控制参数可以根据第一速率和保证时间窗确定，也就是说终端设备可以基于速率分配算力资源。

在一种可能的实现方式中，算力控制参数是一个或多个业务的第二速率和一个或多个业务的第一时长的乘积。其中，第二速率是单位时间内分配的算力大小，第一时长是确定算力控制参数的时间与上一次确定算力控制参数的时间之间的时长。

基于该方案，算力控制参数可以基于第二速率和业务的未分配算力资源的时长来确定，在分配算力资源时可以避免低速率的业务被饿死。

在一种可能的实现方式中，在第二速率和第一时长的乘积大于或等于第一速率和保证时间窗的乘积的情况下，算力控制参数为第一速率和保证时间窗的乘积。在第二速率和第一时长的乘积小于第一速率和保证时间窗的乘积的情况下，算力控制参数为第一速率和第一时长的乘积。

基于该方案，终端设备可以在分配算力资源前更新算力控制参数，从而可以在分配算力资源时，实现较为公平的算力分配，避免高优先级的业务总是被优先分配算力资源始终抢占算力资源，而低优先级的业务总是等待无法获得算力资源，可以避免低优先级业务的业务被饿死。

在一种可能的实现方式中，终端设备基于算力控制参数和一个或多个业务的优先级，为一个或多个业务分配算力资源。

在一种可能的实现方式中，终端设备为一个或多个业务中优先级最高，且算力控制参数大于0的业务分配算力资源。基于该方案，终端设备可以通过优先级和算力控制参数，为业务分配算力资源，从而实现较为公平的算力分配，避免高优先级的业务总是被优先分配算力资源始终抢占算力资源，而低优先级的业务总是等待无法获得算力资源，可以避免低优先级业务的业务被饿死。

在一种可能的实现方式中，终端设备在一个或多个业务中存在速率无穷大的业务的情况下，为速率无穷大的业务分配算力资源。基于该方案，速率无穷大那么可以认为算力控制参数是无穷大的，因此终端设备可以为速率无穷大的也就是算力控制参数无穷大的业务优先分配算力资源。

在一种可能的实现方式中，在存在剩余的算力资源的情况下，丢弃剩余的算力资源。基于该方案，同一个算力资源不能够被不同的业务共享，在存在剩余的算力资源的情况下可以丢弃。

在一种可能的实现方式中，向网络设备发送剩余的算力资源的指示信息，剩余的算力资源的指示

信息用于指示请求去激活剩余的算力资源。基于该方案，终端设备可以将丢弃的算力资源向网络设备指示，从而可以让网络设备将丢弃算力资源重新调度和指示。

在一种可能的实现方式中，在存在剩余的算力资源的情况下，为一个或多个业务中算力控制参数大于0的业务，分配剩余的算力资源。基于该方案，同一个算力资源可以被不同的业务共享，在存在剩余的算力资源的情况下，可以分配给其他的业务。

在一种可能的实现方式中，在存在剩余的算力资源的情况下，按照一个或多个业务中算力控制参数大于0的业务的优先级由高到低的顺序，分配剩余的算力资源。基于该方案，同一个算力资源可以被不同的业务共享，在存在剩余的算力资源的情况下，可以基于业务的优先级和/或算力控制参数进行算力分配。

在一种可能的实现方式中，在存在剩余的算力资源的情况下，且一个或多个业务请求的算力资源大小为0的情况下，为一个或多个业务中算力控制参数小于0的业务分配剩余的算力资源。基于该方案，同一个算力资源可以被不同的业务共享，在存在剩余的算力资源的情况下，虽然算力控制参数小于0但在业务存在算力需求的情况下，也可以为其分配算力资源。

在一种可能的实现方式中，更新算力控制参数，更新的算力控制参数是基于算力控制参数以及为对应业务分配的算力资源确定的。基于该方案，终端设备可以更新算力控制参数，用于下一次分配算力资源。

在一种可能的实现方式中，接收业务与算力资源的关联关系。关联关系用于指示算力资源用于执行业务，或者关联关系用于指示算力资源不可用于执行业务。基于关联关系和/或算力控制参数，为一个或多个业务分配算力资源。基于该方案，终端设备可以为特定的业务预留算力资源。

第二方面，提供一种资源调度方法。该方法可以由终端设备执行，或者芯片/芯片系统执行。该方法中，终端设备激活或去激活第一算力资源，并基于第一算力资源和/或第二算力资源执行第一业务。其中，第一算力资源与第二算力资源不同。这样，可以实现开启或者关闭终端设备的算力资源，以达到减少终端设备能耗或者减少业务的时延的问题。

在一种可能的实现方式中，第一算力资源是辅助算力资源，第二算力资源是主要算力资源。

在一种可能的实现方式中，终端设备基于第一预设条件，激活第一算力资源。其中，第一预设条件包括条件1~条件3中的一项或多项：条件1：根据业务的服务质量（quality of service, QoS）要求，确定激活或者去激活第一算力资源。例如，终端可以预估业务的QoS，确定是否能够满足业务的QoS要求，从而激活或去激活第一算力资源。条件2：根据第二算力资源的负载，确定激活或者去激活第一算力资源。例如，在第二算力资源的负载大于或大于等于第一预设阈值的情况下，终端可以确定需要激活第一算力资源。条件3：基于业务信息与算力资源的关联关系，确定第一算力资源用于执行第一业务或确定第二算力资源不可用于执行第一业务。基于该方案，终端设备可以基于预设的条件确定是否需要激活第一算力资源。

在一种可能的实现方式中，终端设备接收来自网络设备的第一预设条件。基于该方案，终端设备可以基于网络设备配置的条件确定是否激活第一算力资源。

在一种可能的实现方式中，终端设备向网络设备发送第一信息，第一信息包括第一预设条件中满足的一个或多个条件。基于该方案，终端设备可以向网络设备发送终端设备满足的条件，方便网络设备确定是否需要激活第一算力资源。

在一种可能的实现方式中，终端设备基于第二预设条件，去激活第一算力资源。其中，第二预设条件包括条件4~条件6中的一项或多项：

条件4：比较第二算力资源的负载与第二预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况达到第一预设时长的情况下，终端可以确定需要去激活第一算力资源。

条件5：比较第二算力资源的负载与第三预设阈值的大小关系，以及比较第一算力资源与第四预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况达到第二预设时长的情况下，终端可以确定需要去激活第一算力资源。

条件 6: 第一算力资源的开始时长达到第二时长。其中, 第一算力资源的开始时长可以理解为激活、开启或启动第一算力资源的时长。

条件 7: 基于业务信息与算力资源的关联关系, 确定需要去激活第一算力资源。

基于该方案, 终端设备可以基于预设的条件确定是否需要去激活第一算力资源。

5 在一种可能的实现方式中, 终端设备接收来自网络设备的第二预设条件。基于该方案, 终端设备可以基于网络设备配置的条件, 确定是否需要去激活第一算力资源。

在一种可能的实现方式中, 终端设备向网络设备发送第二信息, 第二信息包括第二预设条件中满足的一个或多个条件。基于该方案, 终端设备可以向网络设备发送终端设备满足的条件, 方便网络设备确定是否需要去激活第一算力资源。

10 在一种可能的实现方式中, 终端设备向网络设备发送第三信息和/或第四信息。其中, 第三信息包含第一算力资源的标识, 第三信息指示激活第一算力资源。第四信息包含第一算力资源的标识, 第四信息指示去激活第一算力资源。

基于该方案, 终端设备可以在激活或去激活第一算力资源后, 告知网络设备。

15 在一种可能的实现方式中, 终端设备确定第二算力资源不能满足第一业务的服务质量, 向网络设备发送第三信息, 第三信息指示激活第一算力资源。基于该方案, 终端设备可以基于业务的服务质量确定是否激活第一算力资源。

在一种可能的实现方式中, 终端设备接收来自网络设备的指示信息, 指示信息用于指示激活或去激活第一算力资源。终端设备基于指示信息, 激活或去激活第一算力资源。基于该方案, 终端设备可以基于网络设备的指示信息激活或去激活第一算力资源。

20 在一种可能的实现方式中, 指示信息通过以下中的一项或多项携带: 无线资源控制 (radio resource control, RRC)、媒体接入 (media access control, MAC) 控制元素 (control element, CE) 和下行控制信息 (downlink control information, DCI)。

第三方面, 提供一种资源调度方法。该方法可以由第一装置执行。第一装置可以是终端设备, 或者芯片/芯片系统。该方法中, 第一装置向第二装置发送请求信息, 请求信息用于请求第二装置的算力资源, 请求信息包括计算量、服务质量、业务标识和算力资源标识中的一项或多项。第一装置接收来自第二装置的响应信息, 响应信息指示请求算力资源成功或者请求算力资源失败。

25 可选的, 第一装置可以是调制解调器 (modem), 第二装置可以是应用处理器 (application processing unit, APU)。基于该方案, 调制解调器可以调度 APU 的算力资源来执行业务的计算需求。

30 第四方面, 提供一种资源调度方法。该方法可以由网络设备执行, 或者芯片/芯片系统执行。该方法中, 网络设备确定用于终端设备确定算力控制参数的参数。网络设备向终端设备发送配置信息, 配置信息包括参数。其中, 算力控制参数用于终端设备分配算力资源。

在一种可能的实现方式中, 配置信息包括保证时间窗和速率, 算力控制参数是速率和保证时间窗的乘积, 速率是单位时间内分配的算力大小, 保证时间窗指用于确定算力控制参数的时长。

35 在一种可能的实现方式中, 网络设备向终端设备发送业务与算力资源的关联关系。关联关系用于指示算力资源用于执行业务, 或者关联关系用于指示算力资源不可用于执行业务。

第五方面, 提供一种资源调度方法。该方法可以由网络设备执行, 或者芯片/芯片系统执行。该方法中, 网络设备向终端设备发送第一预设条件和/或第二预设条件。第一预设条件用于终端设备确定是否激活第一算力资源, 第二预设条件用于终端设备确定是否去激活第一算力资源。

40 其中, 第一预设条件包括条件 1~条件 3 中的一项或多项: 条件 1: 根据业务的服务质量 (quality of service, QoS) 要求, 确定激活或者去激活第一算力资源。例如, 终端可以预估业务的 QoS, 确定是否能够满足业务的 QoS 要求, 从而激活或去激活第一算力资源。条件 2: 根据第二算力资源的负载, 确定激活或者去激活第一算力资源。例如, 在第二算力资源的负载大于或大于等于第一预设阈值的情况下, 终端可以确定需要激活第一算力资源。条件 3: 基于业务信息与算力资源的关联关系, 确定第一算力资源用于执行第一业务或确定第二算力资源不可用于执行第一业务。基于该方案, 终端设备可以基于预设的条件确定是否需要激活第一算力资源。

45 其中, 第二预设条件包括条件 4~条件 7 中的一项或多项:

条件 4: 比较第二算力资源的负载与第二预设阈值的大小关系, 确定需要去激活第一算力资源。例如, 在第二算力资源的负载小于或小于等于第二预设阈值的情况下, 终端可以确定需要去激活第一算力

资源。又例如,在第二算力资源的负载小于或小于等于第二预设阈值的情况达到第一预设时长的情况下,终端可以确定需要去激活第一算力资源。

条件 5: 比较第二算力资源的负载与第三预设阈值的大小关系,以及比较第一算力资源与第四预设阈值的大小关系,确定需要去激活第一算力资源。例如,在第二算力资源的负载小于或小于等于第三预设阈值,且第一算力资源的负载小于或小于等于第四预设阈值的情况下,终端可以确定需要去激活第一算力资源。又例如,在第二算力资源的负载小于或小于等于第三预设阈值,且第一算力资源的负载小于或小于等于第四预设阈值的情况达到第二预设时长的情况下,终端可以确定需要去激活第一算力资源。

条件 6: 第一算力资源的开始时长达到第二时长。其中,第一算力资源的开始时长可以理解为激活、开启或启动第一算力资源的时长。

条件 7: 基于业务信息与算力资源的关联关系,确定需要去激活第一算力资源。

在一种可能的实现方式中,网络设备接收来自终端设备的第三信息和/或第四信息。其中,第三信息包含第一算力资源的标识,第三信息指示激活第一算力资源。第四信息包含第一算力资源的标识,第四信息指示去激活第一算力资源。

在一种可能的实现方式中,网络设备向终端设备发送指示信息,指示信息用于指示激活或去激活第一算力资源。

在一种可能的实现方式中,指示信息通过以下中的一项或多项携带: RRC、MAC CE 和 DCI。

第六方面,提供一种资源调度方法。该方法可以由第二装置执行。第二装置可以是终端设备,或者芯片/芯片系统。该方法中,第二装置接收来自第一装置的请求信息,请求信息用于请求第二装置的算力资源,请求信息包括计算量、服务质量、业务标识和算力资源标识中的一项或多项。第二装置向第一装置发送响应信息,响应信息指示请求算力资源成功或者请求算力资源失败。

第七方面,提供一种通信装置,包括处理单元。

处理单元,用于获取算力资源。处理单元还用于,确定一个或多个业务的算力控制参数,并基于算力控制参数为一个或多个业务分配算力资源。

在一种可能的实现方式中,所述通信装置还包括:收发单元,用于接收来自网络设备的配置信息,该配置信息包括用于确定算力控制参数的参数。

在一种可能的实现方式中,配置信息包括保证时间窗和速率,算力控制参数是速率和保证时间窗的乘积,速率是单位时间内分配的算力大小,保证时间窗指用于确定算力控制参数的时长。

在一种可能的实现方式中,算力控制参数是一个或多个业务的速率和一个或多个业务的第一时长的乘积。其中,第一时长是确定算力控制参数的时间与上一次确定算力控制参数的时间之间的时长。

在一种可能的实现方式中,在速率和第一时长的乘积大于或等于速率和保证时间窗的乘积的情况下,算力控制参数为速率和保证时间窗的乘积。在速率和第一时长的乘积小于速率和保证时间窗的乘积的情况下,算力控制参数为速率和第一时长的乘积。其中,保证时间窗指用于确定算力控制参数的时长。

在一种可能的实现方式中,处理单元,还用于基于算力控制参数和一个或多个业务的优先级,为一个或多个业务分配算力资源。

在一种可能的实现方式中,处理单元,还用于为一个或多个业务中优先级最高,且算力控制参数大于 0 的业务分配算力资源。

在一种可能的实现方式中,处理单元,还用于在一个或多个业务中存在速率无穷大的业务的情况下,为速率无穷大的业务分配算力资源。

在一种可能的实现方式中,处理单元,还用于在存在剩余的算力资源的情况下,丢弃剩余的算力资源。

在一种可能的实现方式中,收发单元,还用于向网络设备发送剩余的算力资源的指示信息,剩余的算力资源的指示信息用于指示请求去激活剩余的算力资源。

在一种可能的实现方式中,处理单元,还用于在存在剩余的算力资源的情况下,为一个或多个业务中算力控制参数大于 0 的业务,分配剩余的算力资源。

在一种可能的实现方式中,处理单元,还用于在存在剩余的算力资源的情况下,按照一个或多个业务中算力控制参数大于 0 的业务的优先级由高到低的顺序,分配剩余的算力资源。

在一种可能的实现方式中,处理单元,还用于在存在剩余的算力资源的情况下,且一个或多个业务请求的算力资源大小为 0 的情况下,为一个或多个业务中算力控制参数小于 0 的业务分配剩余的算力资源。

资源。

在一种可能的实现方式中，处理单元，还用于更新算力控制参数，更新的算力控制参数是基于算力控制参数以及为对应业务分配的算力资源确定的。

5 在一种可能的实现方式中，收发单元，还用于接收业务与算力资源的关联关系。关联关系用于指示算力资源用于执行业务，或者关联关系用于指示算力资源不可用于执行业务。处理单元，还用于基于关联关系和/或算力控制参数，为一个或多个业务分配算力资源。

第八方面，提供一种通信装置，包括处理单元。

处理单元，用于激活或去激活第一算力资源，并基于第一算力资源和/或第二算力资源执行第一业务。其中，第一算力资源与第二算力资源不同。

10 在一种可能的实现方式中，第一算力资源是辅助算力资源，第二算力资源是主要算力资源。

在一种可能的实现方式中，处理单元，还用于基于第一预设条件，激活第一算力资源。其中，第一预设条件包括条件1~条件3中的一项或多项：条件1：根据业务的服务质量（quality of service, QoS）要求，确定激活或者去激活第一算力资源。例如，终端可以预估业务的QoS，确定是否能够满足业务的QoS要求，从而激活或去激活第一算力资源。条件2：根据第二算力资源的负载，确定激活或者去激活第一算力资源。例如，在第二算力资源的负载大于或大于等于第一预设阈值的情况下，终端可以确定需要激活第一算力资源。条件3：基于业务信息与算力资源的关联关系，确定第一算力资源用于执行第一业务或确定第二算力资源不可用于执行第一业务。基于该方案，终端设备可以基于预设的条件确定是否需要激活第一算力资源。

20 在一种可能的实现方式中，所述通信装置还包括：收发单元，还用于接收来自网络设备的第一预设条件。

在一种可能的实现方式中，收发单元，还用于向网络设备发送第一信息，第一信息包括第一预设条件中满足的一个或多个条件。

在一种可能的实现方式中，处理单元，还用于基于第二预设条件，去激活第一算力资源。其中，第二预设条件包括条件4~条件6中的一项或多项：

25 条件4：比较第二算力资源的负载与第二预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况达到第一预设时长的情况下，终端可以确定需要去激活第一算力资源。

30 条件5：比较第二算力资源的负载与第三预设阈值的大小关系，以及比较第一算力资源与第四预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况达到第二预设时长的情况下，终端可以确定需要去激活第一算力资源。

35 条件6：第一算力资源的开始时长达到第二时长。其中，第一算力资源的开始时长可以理解为激活、开启或启动第一算力资源的时长。

条件7：基于业务信息与算力资源的关联关系，确定需要去激活第一算力资源。

在一种可能的实现方式中，收发单元，还用于接收来自网络设备的第二预设条件。基于该方案，终端设备可以基于网络设备配置的条件，确定是否需要去激活第一算力资源。

40 在一种可能的实现方式中，收发单元，还用于向网络设备发送第二信息，第二信息包括第二预设条件中满足的一个或多个条件。

在一种可能的实现方式中，收发单元，还用于向网络设备发送第三信息和/或第四信息。其中，第三信息包含第一算力资源的标识，第三信息指示激活第一算力资源。第四信息包含第一算力资源的标识，第四信息指示去激活第一算力资源。

45 在一种可能的实现方式中，处理单元，还用于确定第二算力资源不能满足第一业务的服务质量，向网络设备发送第三信息，第三信息指示激活第一算力资源。

在一种可能的实现方式中，收发单元，还用于接收来自网络设备的指示信息，指示信息用于指示激活或去激活第一算力资源。处理单元，还用于基于指示信息，激活或去激活第一算力资源。

在一种可能的实现方式中，指示信息通过以下中的一项或多项携带：RRC、MAC CE和DCI。

第九方面，提供一种通信装置，包括：处理单元和收发单元。

处理单元，用于生成请求信息。收发单元，用于向第二装置发送请求信息，请求信息用于请求所第二装置的算力资源，请求信息包括计算量、服务质量、业务标识和算力资源标识中的一项或多项。收发单元，还用于接收来自第二装置的响应信息，响应信息指示请求算力资源成功或者请求算力资源失败。

5 第十方面，提供一种通信装置，包括：处理单元和收发单元。

处理单元，用于确定用于终端设备确定算力控制参数的参数。收发单元，用于向终端设备发送配置信息，配置信息包括参数。其中，算力控制参数用于终端设备分配算力资源。

在一种可能的实现方式中，配置信息包括保证时间窗和速率，算力控制参数是速率和保证时间窗的乘积，速率是单位时间内分配的算力大小，保证时间窗指用于确定算力控制参数的时长。

10 在一种可能的实现方式中，收发单元，还用于向终端设备发送业务与算力资源的关联关系。关联关系用于指示算力资源用于执行业务，或者关联关系用于指示算力资源不可用于执行业务。

第十一方面，提供一种通信装置，包括：处理单元和收发单元。

处理单元，用于确定第一预设条件和/或第二预设条件。收发单元，用于向终端设备发送第一预设条件和/或第二预设条件。第一预设条件用于终端设备确定是否激活第一算力资源，第二预设条件用于终端设备确定是否去激活第一算力资源。

15 其中，第一预设条件包括条件1~条件3中的一项或多项：条件1：根据业务的服务质量（quality of service, QoS）要求，确定激活或者去激活第一算力资源。例如，终端可以预估业务的QoS，确定是否能够满足业务的QoS要求，从而激活或去激活第一算力资源。条件2：根据第二算力资源的负载，确定激活或者去激活第一算力资源。例如，在第二算力资源的负载大于或大于等于第一预设阈值的情况下，终端可以确定需要激活第一算力资源。条件3：基于业务信息与算力资源的关联关系，确定第一算力资源用于执行第一业务或确定第二算力资源不可用于执行第一业务。基于该方案，终端设备可以基于预设的条件确定是否需要激活第一算力资源。

其中，第二预设条件包括条件4~条件7中的一项或多项：

25 条件4：比较第二算力资源的负载与第二预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第二预设阈值的情况达到第一预设时长的情况下，终端可以确定需要去激活第一算力资源。

条件5：比较第二算力资源的负载与第三预设阈值的大小关系，以及比较第一算力资源与第四预设阈值的大小关系，确定需要去激活第一算力资源。例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况下，终端可以确定需要去激活第一算力资源。又例如，在第二算力资源的负载小于或小于等于第三预设阈值，且第一算力资源的负载小于或小于等于第四预设阈值的情况达到第二预设时长的情况下，终端可以确定需要去激活第一算力资源。

30 条件6：第一算力资源的开始时长达到第二时长。其中，第一算力资源的开始时长可以理解为激活、开启或启动第一算力资源的时长。

35 条件7：基于业务信息与算力资源的关联关系，确定需要去激活第一算力资源。

在一种可能的实现方式中，收发单元，还用于接收来自终端设备的第三信息和/或第四信息。其中，第三信息包含第一算力资源的标识，第三信息指示激活第一算力资源。第四信息包含第一算力资源的标识，第四信息指示去激活第一算力资源。

40 在一种可能的实现方式中，收发单元，还用于向终端设备发送指示信息，指示信息用于指示激活或去激活第一算力资源。

在一种可能的实现方式中，指示信息通过以下中的一项或多项携带：RRC、MAC CE和DCI。

第十二方面，提供一种通信装置，包括：处理单元和收发单元。

收发单元，用于接收来自第一装置的请求信息，请求信息用于请求所第二装置的算力资源，请求信息包括计算量、服务质量、业务标识和算力资源标识中的一项或多项。处理单元，用于生成响应信息。收发单元，还用于向第一装置发送响应信息，响应信息指示请求算力资源成功或者请求算力资源失败。

45 第十三方面，本申请实施例提供一种通信装置，该通信装置可以为上述实施例中第七方面至第十二方面中任一方面的通信装置，或者为设置在第七方面至第十二方面中任一方面的通信装置中的芯片。该通信装置包括通信接口以及处理器，可选的，还包括存储器。其中，该存储器用于存储计算机程序或指

令或者数据,处理器与存储器、通信接口耦合,当处理器读取所述计算机程序或指令或数据时,使通信装置执行上述第一方面至第六方面中任一方面方法实施例中网络设备或终端设备所执行的方法。

应理解,该通信接口可以通过所述通信装置中的天线、馈线和编解码器等实现,或者,如果通信装置为设置在通信设备中的芯片,则通信接口可以是该芯片的输入/输出接口,例如输入/输出管脚等。所述通信装置还可以包括收发器,用于该通信装置与其它设备进行通信。

第十四方面,本申请提供了一种通信装置,包括:逻辑电路和输入输出接口。逻辑电路用于执行上述第一方面至第六方面中任一种可能的实现方式中网络设备或终端设备所执行的方法。输入输出接口用于与其他通信装置通信。

第十五方面,本申请实施例提供了一种芯片系统,该芯片系统包括处理器,还可以包括存储器,用于实现第一方面至第六方面中任一方面的网络设备或终端设备执行的方法。在一种可能的实现方式中,所述芯片系统还包括存储器,用于保存程序指令和/或数据。该芯片系统可以由芯片构成,也可以包含芯片和其他分立器件。

第十六方面,本申请实施例提供了一种通信系统,所述通信系统包括第七方面和第十方面中任一方面的通信装置。或者,所述通信系统包括第八方面和第十一方面中任一方面的通信装置。或者,所述通信系统包括第九方面和第十二方面中任一方面的通信装置。

第十七方面,本申请提供了一种计算机可读存储介质,该计算机可读存储介质存储有计算机程序或指令,当该计算机程序或指令被运行时,实现上述各方面中由网络设备或终端设备执行的方法。

第十八方面,提供了一种计算机程序产品,所述计算机程序产品包括:计算机程序代码或指令,当所述计算机程序代码或指令被运行时,使得上述各方面中由网络设备或终端设备执行的方法被执行。

第十九方面,提供了一种通信装置,所述通信装置包括执行上述各方面方法的单元或模块。

上述第四方面至第十九方面及其实现方式的有益效果可以参考对第一方面至第三方面的方法及其实现方式的有益效果的描述。

附图说明

- 图1为本申请实施例提供的通信系统的示意图;
图2为本申请实施例提供的一种资源调度方法的示例性流程图;
图3为本申请实施例提供的一种令牌桶的示意图;
图4为本申请实施例提供的一种白名单的示意图;
图5为本申请实施例提供的又一种资源调度方法的示例性流程图;
图6为本申请实施例提供的又一种资源调度方法的示例性流程图;
图7为本申请实施例提供的一种API的示意图;
图8为本申请实施例提供的一种通信装置的示意图;
图9为本申请实施例提供的又一种通信装置的示意图;
图10为本申请实施例提供的又一种通信装置的示意图;
图11为本申请实施例提供的又一种通信装置的示意图。

具体实施方式

为了便于理解本申请实施例提供的技术方案,以下对本申请涉及的技术术语进行解释和说明。

1) 算力,指处理信息的能力,也指执行某种计算需求的能力。

2) 算力资源,用于执行计算需求的资源,如中央处理器(central processing unit, CPU)或图形处理器(graphics processing unit, GPU)等。可选的,从时域上看,时间片(time slot)为算力的最小资源。

3) 业务,又可以称为任务、task、或者应用(application, APP)等。

4) 激活某个算力资源,又可以称为开启某个算力资源或启动某个算力资源,指采用某个算力资源来执行计算需求。类似的,去激活某个算力资源,又可以称为关闭某个算力资源,指不采用某个算力资源来执行计算需求。

下面结合说明书附图对本申请实施例做详细描述。

本申请实施例的技术方案可以应用于新无线(New Radio, NR)系统、长期演进(Long Term Evolution,

LTE)系统、LTE 频分双工(Frequency Division Duplex, FDD)系统、LTE 时分双工(Time Division Duplex, TDD)、全球互联微波接入(Worldwide Interoperability for Microwave Access, WiMAX)通信系统等,在此不做限制。

图 1 是本申请的实施例应用的通信系统 1000 的架构示意图。如图 1 所示,该通信系统包括无线接入网 100。其中,无线接入网 100 可以包括至少一个网络设备(如图 1 中的 110a 和/或 110b),还可以包括至少一个终端装置(如图 1 中的 120a-120j 中的至少一个)。终端装置通过无线的方式与接入网设备相连,接入网设备通过无线或有线方式与核心网设备连接。终端装置和终端装置之间以及网络设备和网络设备之间可以通过有线或无线的方式相互连接。图 1 只是示意图,该通信系统中还可以包括其它网络设备,如还可以包括无线中继设备和无线回传设备,在图 1 中未画出。

网络设备是一种具有无线收发功能的网络侧设备。网络设备可以是无线接入网(radio access network, RAN)中为终端设备提供无线通信功能的装置,称为 RAN 设备。例如,网络设备可以是基站(base station)、演进型基站(evolved NodeB, eNB)、发送接收点(transmission reception point, TRP)、第五代(5th generation, 5G)移动通信系统中的下一代基站(next generation NodeB, gNB)、第六代(6th generation, 6G)移动通信系统中的下一代基站、未来移动通信系统中的基站或 Wi-Fi 系统中的接入节点等;也可以是完成基站部分功能的模块或单元,例如,可以是集中式单元(central unit, CU),也可以是分布式单元(distributed unit, DU)。这里的 CU 完成基站的无线资源控制协议和分组数据汇聚层协议(packet data convergence protocol, PDCP)的功能,还可以完成业务数据适配协议(service data adaptation protocol, SDAP)的功能;DU 完成基站的无线链路控制层和介质访问控制(media access control, MAC)层的功能,还可以完成部分物理层或全部物理层的功能,有关上述各个协议层的具体描述,可以参考第三代合作伙伴计划(3rd generation partnership project, 3GPP)的相关技术规范。网络设备可以是宏基站(如图 1 中的 110a),也可以是微基站或室内站(如图 1 中的 110b),还可以是中继节点或施主节点等。本申请的实施例对网络设备所采用的具体技术和具体设备形态不做限定。本申请实施例中以基站为网络设备为例进行说明。

在另一种可能的场景中,由多个 RAN 节点协作协助终端实现无线接入,不同 RAN 节点分别实现基站的部分功能。例如,RAN 节点可以是 CU, DU, CU-控制面(control plane, CP), CU-用户面(user plane, UP),或者无线单元(radio unit, RU)等。CU 和 DU 可以是单独设置,或者也可以包括在同一个网元中,例如基带单元(baseband unit, BBU)中。RU 可以包括在射频设备或者射频单元中,例如包括在射频拉远单元(remote radio unit, RRU)、有源天线处理单元(active antenna unit, AAU)或远程射频头(remote radio head, RRH)中。

在不同系统中,CU(或 CU-CP 和 CU-UP)、DU 或 RU 也可以有不同的名称,但是本领域的技术人员可以理解其含义。例如,在 ORAN 系统中,CU 也可以称为 O-CU(开放式 CU),DU 也可以称为 O-DU, CU-CP 也可以称为 O-CU-CP, CU-UP 也可以称为 O-CU-UP, RU 也可以称为 O-RU。为描述方便,本申请中以 CU, CU-CP, CU-UP, DU 和 RU 为例进行描述。本申请中的 CU(或 CU-CP, CU-UP)、DU 和 RU 中的任一单元,可以通过软件模块、硬件模块、或者软件模块与硬件模块结合来实现。

终端设备是一种具有无线收发功能的用户侧设备。终端设备也可以称为用户设备(user equipment, UE)、移动台、移动终端等。终端装置可以广泛应用于各种场景,例如,设备到设备(device-to-device, D2D)、车物(vehicle to everything, V2X)通信、机器类通信(machine-type communication, MTC)、物联网(internet of things, IOT)、虚拟现实、增强现实、工业控制、自动驾驶、远程医疗、智能电网、智能家居、智能办公、智能穿戴、智能交通、智慧城市等。终端装置可以是手机、平板电脑、带无线收发功能的电脑、可穿戴设备、车辆、无人机、直升机、飞机、轮船、机器人、机械臂、智能家居设备等。本申请的实施例对终端装置所采用的具体技术和具体装置形态不做限定。本申请实施例以终端为终端设备为例进行说明。

网络设备和终端设备可以是固定位置的,也可以是可移动的。网络设备和终端设备可以部署在陆地上,包括室内或室外、手持或车载;也可以部署在水面上;还可以部署在空中的飞机、气球和人造卫星上。本申请的实施例对网络设备和终端设备的应用场景不做限定。

网络设备和终端设备的角色可以是相对的,例如,图 1 中的直升机或无人机 120i 可以被配置成移动网络设备,对于那些通过 120i 接入到无线接入网 100 的终端设备 120j 来说,终端设备 120i 是网络设备;但对于网络设备 110a 来说,120i 是终端设备,即 110a 与 120i 之间是通过无线空口协议进行通信

的。当然，110a 与 120i 之间也可以是通过网络设备与网络设备之间的接口协议进行通信的，此时，相对于 110a 来说，120i 也是网络设备。因此，网络设备和终端设备都可以统一称为通信装置，图 1 中的 110a 和 110b 可以称为具有网络设备功能的通信装置，图 1 中的 120a-120j 可以称为具有终端设备功能的通信装置。

在本申请的实施例中，网络设备的功能也可以由网络设备中的模块（如芯片）来执行，也可以由包含有网络设备功能的控制子系统来执行。这里的包含有网络设备功能的控制子系统可以是智能电网、工业控制、智能交通、智慧城市等上述应用场景中的控制中心。终端设备的功能也可以由终端设备中的模块（如芯片或调制解调器）来执行，也可以由包含有终端设备功能的装置来执行。

伴随着人工智能（artificial intelligence, AI）超大模型，以及扩展现实（extended reality, XR）类等大计算量业务的出现，无论是网络设备或是终端设备，单点计算瓶颈都会逐步显现。因此在无线网络中支持算力共享是未来趋势。随着泛终端（如对于支持自动驾驶的以及具有不同等级的算力能力的车终端）的引入，终端设备的能力不断增强，终端可共享和利用其算力给网络从而辅助网络计算。目前尚无如何共享终端设备算力的技术方案。在云计算中，已支持分布式计算对应的算力共享和算力调度等功能，但一般都是在云网络中、基于固定网络实施和部署。例如，针对多个计算节点，由统一的集中管控节点来集中管理和调度各个计算节点。但是，云网络支持简单的算力共享和调度，其基于固定网络的方法无法重用至无线网络中。不仅如此，相关技术中也没有终端在获取到算力资源后如何为任务分配算力的技术方案。

鉴于此，本申请实施例提供一种资源调度方法。该方法中，终端设备可以获取算力资源，并确定一个或多个业务的算力控制参数。终端设备可以基于算力控制参数为一个或多个业务分配获取的算力资源。基于该方案，终端设备可以基于算力控制参数为有计算需求的业务分配算力资源，实现终端设备的算力的共享。

参阅图 2，为本申请实施例提供的一种资源调度方法的示例性流程图，可以包括以下操作。图 2 所示的实施例可以由图 1 所示的终端设备和网络设备执行。下文中，以终端设备为终端，网络设备为基站为例进行说明。

可选的，图 2 所示的实施例可以包括以下操作 S201。

S201：终端获取算力资源。

一种可能的情况中，终端可以获取预配置的算力资源。

另一种可能的情况中，终端可以向基站请求算力资源，由基站为终端进行算力资源的调度或分配。举例来说，终端可以接收来自基站的算力资源指示信息。该算力资源指示信息可以指示算力资源，终端可以获取算力资源指示信息指示的算力资源。

S202：终端确定一个或多个业务的算力控制参数。

例如，终端可以获取一个或多个业务的优先级（task priority）作为算力控制参数。又例如，终端可以获取一个或多个业务的令牌桶作为算力控制参数。又例如，终端可以将业务的优先级和令牌桶结合作为算力控制参数。可选的，算力控制参数可以用于表征一个或多个业务的算力分配的优先级。

在一种可能的实现方式中，终端可以基于基站的配置信息，确定算力控制参数。基站可以向终端发送配置信息，该配置信息可以包括用于确定算力控制参数的参数。例如，配置信息可以包括一个或多个业务的优先级，终端可以将优先级作为业务的算力控制参数。又例如，配置信息可以包括用于确定业务的令牌桶的参数，如配置信息可以包括保证时间窗和速率。终端可以将保证时间窗和速率的乘积作为业务的令牌桶，也就是业务的算力控制参数。其中，速率是单位时间内分配的算力大小，如可以是 CPU/GPU 时间片、CPU/GPU 占比或者计算量等。保证时间窗（BSD）指用于确定算力控制参数的时长。

需要说明的是，本文中的速率还可以称为优先算力比（prioritized computing rate, PCR）表示，也可以采用其他名称，本申请不做具体限定。

另一种可能的实现方式中，终端可以自主确定并维护业务的算力控制参数，如令牌桶。参阅图 3，终端可以维护各个业务的令牌桶。其中，终端可以令令牌桶的总蓄水量 $c_j = c_{j-1} + \text{PCR} \times T$ （第一时长）。其中，第一时长可以是确定 c_j 与确定 c_{j-1} 时间之间的时长，速率可以是基站指示的或者预配置的。

可以理解的是，下标 j 可以理解为第 j 次分配算力资源时，令牌桶内的总蓄水量，如 c_j 可以理解为第 j 次分配算力资源时，令牌桶内的总蓄水量， c_{j-1} 可以理解为第 $j-1$ 次分配算力资源时，令牌桶内的总

蓄水量, j 可以是大于或等于 0 的整数。本文中, 以 j 的初始值为 0 为例进行说明, 可以理解的是 j 的初始值还可以是 1、2 等预先设定的值。可选的, c_j 可以具有初始值, 该初始值可以是预先设置的, 也就是 c_0 可以是预先设置的, 如可以设置为 0, 或者可以设置为与优先级有关的值, 也及时 c_0 可以设置为与优先级有关的值, 如优先级越高初始值越高。需要说明的是, 本文中 c_j 可以大于 0、等于 0 也可以小于 0。

可选的, 终端可以结合上述两种可能的实现方式确定业务的算力控制参数, 如令牌桶。举例来说, 终端可以自行确定业务的令牌桶的总蓄水量 c_j , 如令 $c_j = c_{j-1} + \text{PCR}(\text{速率}) * T(\text{第一时长})$ 。终端在进行算力分配前, 如执行 S203 前可以基于站的配置信息更新令牌桶的总蓄水量 c_j 。例如, 终端可以确定 c_j 与 $\text{PCR} * \text{BSD}$ 的大小, 从而更新令牌桶。举例来说, 在 c_j 大于或者大于等于 $\text{PCR} * \text{BSD}$ 的情况下, 终端可以将 $\text{PCR} * \text{BSD}$ 作为业务的总蓄水量 c_j 。在 c_j 小于或者小于或等于 $\text{PCR} * \text{BSD}$ 的情况下, 仍将 c_j 作为业务的总蓄水量 c_j 。这样, 终端可以为各个业务的令牌桶注水, 并通过令牌桶内的总蓄水量为业务分配算力资源, 从而在分配算力资源时, 可以考虑到业务的速率, 也可以避免低速率的业务被饿死。

S203: 终端基于算力控制参数为一个或多个业务分配算力资源。

例如, 终端可以按照算力控制参数由高到低的顺序为一个或多个业务分配算力资源。比如, 算力控制参数为优先级的情况下, 终端可以按照优先级由高到低的顺序为一个或多个业务分配算力资源。基于该方案, 终端可以为高优先级业务优先分配算力资源。

又比如, 算力控制参数为令牌桶的情况下, 终端可以按照令牌桶内的总蓄水量由高到低的顺序为一个或多个业务分配算力资源。可选的, 在算力控制参数包括优先级和令牌桶的情况下, 终端可以结合令牌桶和优先级为一个或多个业务分配算力资源。例如, 终端可以根据优先级由高到低的顺序, 且令牌桶内的总蓄水量大于 0 的业务优先分配算力资源。又例如, 终端可以按照令牌桶内的总蓄水量由高到低的顺序为一个或多个业务分配算力资源, 在总蓄水量相同的情况下终端可以为高优先级的业务优先分配算力资源。可选的, 终端可以为速率无穷大 (infinity) 的业务优先分配算力资源, 因为按照 S202 中描述的确定令牌桶的方式, 速率无穷大那么可以认为 c_j 是无穷大的, 因此终端可以为速率无穷大的也就是 c_j 无穷大的业务优先分配算力资源。

基于该方案, 终端可以通过令牌桶内的总蓄水量 c_j 的大小, 为业务分配算力资源, 从而实现较为公平的算力分配, 从而避免高优先级的业务总是被优先分配算力资源始终抢占算力资源, 而低优先级的业务总是等待无法获得算力资源, 可以避免低优先级业务的业务被饿死。

在一种可能的实现方式中, 针对某个算力资源, 可以允许不同的业务复用也可以不允许不同的业务复用。举例来说, 终端为业务 1 分配了算力资源 A, 如时间片 n (n 为大于或等于 1 的整数), 而算力资源 A 在满足业务 1 的计算需求的情况下, 仍存在剩余的算力资源, 也就是还可以执行其他的计算需求, 该剩余的算力资源可以被分配给其他的业务, 如业务 2、业务 3 等, 或者该剩余的算力资源不可以被分配给其他的业务, 需要被丢弃。

可选的, 在剩余的算力资源不可以被分配给其他的业务的情况下, 终端可以向基站发送剩余的算力资源的指示信息, 如剩余的算力资源的标识或者索引等, 从而指示去激活该剩余的算力资源。

一种可能的情况中, 如果某个算力资源允许不同的业务复用, 那么终端可以按照 S203 所描述的方式, 将该算力资源所剩余的资源分配给其他的业务, 直至无剩余的算力资源或者直至不存在有计算需求的业务。例如, 终端可以按照优先级由高到低的顺序, 且 $c_j > 0$ 的有计算需求的业务分配该剩余的算力资源。

可选的, 在为 $c_j > 0$ 的有计算需求的业务分配算力资源之后, 如果仍有剩余的算力资源, 终端可以为有算力需求的业务, 分配该剩余的算力资源, 直至无剩余的算力资源或者直至不存在有算力需求的业务。可以理解的是, 上述有算力需求的业务 c_j 可能小于 0, 如果有为 $c_j > 0$ 的有计算需求的业务分配算力资源之后还有剩余的算力资源, 那么即使 $c_j < 0$ 也可以为该业务分配剩余的算力资源。

另一种可能的情况中, 如果某个算力资源不允许不同的业务复用, 那么终端可以将该算力资源所剩余的资源丢弃。可以理解的是, 丢弃算力资源可以理解为不使用该算力资源执行任何的计算的任务。可选的, 终端可以向基站发送丢弃的算力资源的标识, 如时间片。举例来说, 时间片 $n1 \sim n3$ 中时间片 $n1 \sim n2$ 被分配给业务 1, $n1 \leq n2 < n3$, 而时间片 $n2 \sim n3$ 是剩余的算力资源, 由于不允许不同的业务复用一个算力资源, 因此终端需要丢弃时间片 $n2 \sim n3$, 并向基站发送时间片 $n2 \sim n3$ 的标识。

另一种可能的实现方式中, 终端可以为特定的业务预留算力资源。例如, 终端可以获取白名单或者

黑名单。其中，白名单可以包含业务信息与算力资源的关联关系，白名单可以指示算力资源可以用于执行具有关联关系的业务信息对应的业务的计算需求。黑名单也可以包含业务信息与算力资源的关联关系，黑名单可以指示算力不可用于执行具有关联关系的业务信息对应的业务的计算需求。可选的，黑名单或者白名单可以是基站指示的或者预配置的。

5 需要说明的是，上述业务信息可以包括业务标识和/或逻辑信道。其中，逻辑信道可以与业务一一对应，或者多个业务可以对应一个逻辑信道，本申请不做具体限定。

终端为特定的业务预留的算力资源可以是单次，或者多次。其中，多次预留的算力资源可以是周期性的或者非周期性的，本申请不做具体限定。可选的，终端为特定的业务预留的算力资源的信息可以是白名单或者黑名单中指示的。

10 参阅图 4，假设白名单中 CPU 中的用黑色矩形标注的时间片 M 是与业务 1 关联的。那么，终端可以确定业务 1 是否具有算力需求，如果业务 1 具有算力需求，那么终端可以将时间片 M 分配给业务 1。如果业务 1 没有算力需求，那么终端可以丢弃时间片 M。可以理解的是，由于时间片 M 是与业务 1 关联的，因此即使时间片 M 有剩余资源，也不可以分配给其他业务。如果具有算力需求的其他业务，如业务 2~业务 N，终端可以为业务 2~业务 N 分配图 4 中未被黑色矩形标注的时间片 X。可以理解的是，
15 终端为业务 2~业务 N 分配算力资源，如时间片的方式可以参照前述 S203 实施，此处不再赘述。

在执行 S203 之前，终端可以将按照优先级由高到低的顺序对业务进行排序，并根据业务的 c_j 的大小分配算力资源，可以参照前述相关描述，此处不再赘述。

可选的，终端可以根据前述白名单或者黑名单对算力资源进行过滤，也就是确定一个或多个业务可以被分配哪些算力资源，并根据业务的 c_j 的大小分配算力资源。例如，终端根据白名单或者黑名单，
20 确定算力资源 X 仅可用于业务 A，那么终端可以将算力资源 X 分配给业务 A。并将其余的算力资源，如算力资源 M 按照 S203 的描述分配给其余有计算需求的业务。又例如，终端根据白名单或者黑名单，确定算力资源 X 可用于业务 A 和业务 B，那么终端按照 S203 的描述将算力资源 X 分配给业务 A 和/或业务 B。并将其余的算力资源，如算力资源 M 按照 S203 的描述分配给其余有计算需求的业务。可以理解的是，有关资源复用的方式可以参照前述相关描述，此处不再赘述。

25 在一个示例中，参阅图 3，终端在执行 S203 后，可以更新业务的算力控制参数，如更新业务的令牌桶。其中，令牌桶内的总蓄水量 $c_j = c_j - \text{分配的算力资源的大小}$ 。可选的，分配的算力资源的大小可以以时间片占用的时长表示，如 ms、s 等。举例来说，针对业务 1，终端为业务 1 分配了时间片 1、2 和 3，时间片 1、2 和 3 共占用 10s，那么终端可以更新业务 1 的令牌桶，令业务 1 的令牌桶内的总蓄水量 $c_j = c_j - 10$ 。

30 基于上述方案，终端可以基于令牌桶内的注水和消耗的过程，为各个业务分配算力资源，从而可以解决多业务场景下的算力分配的问题，可以避免低优先级或者低速率的业务被饿死。

本申请实施例中，基于上述实施例介绍了在多业务场景下，终端分配算力资源的方法。目前，终端可以具有多个 CPU/GPU，也就是终端可以是多核的，如何控制终端开启或者关闭多个 CPU，以减少终端能耗或者减少业务的时延是一个需要解决的技术问题。

35 鉴于此，本申请实施例还提供另一种资源调度方法。该方法中，终端可以自行确定激活或者去激活第一算力资源，或者基于基站的指示激活或去激活第一算力资源。这样，可以实现开启或者关闭终端的 CPU，以达到减少终端能耗或者减少业务的时延的有益效果。

参阅图 5，为本申请实施例提供的一种资源调度方法的示例性流程图，可以包括以下操作。可以理解的是，图 5 所示的实施例可以由图 1 所示的终端设备和网络设备执行。下文中，以终端设备为终端，
40 网络设备为基站为例进行说明。

S501: 终端激活或去激活第一算力资源。

例如，终端可以根据预先配置的条件，自行确定激活或去激活第一算力资源。又例如，终端可以基于基站的指示，激活或去激活第一算力资源。

可选的，图 5 所示的实施例还可以包括以下操作 S502。

45 S502: 终端基于第一算力资源和/或第二算力资源执行第一业务。

例如，终端可以基于第一算力资源和/或第二算力资源来执行第一业务的计算的任务。其中，第一算力资源和第二算力资源是不同的算力资源。例如，第一算力资源可以是辅助算力资源，如辅 CPU/辅 GPU，第二算力资源可以是主要（primary）算力资源，如主 CPU/主 GPU。

以下,通过方式一和方式二介绍 S501 中终端激活或去激活第一算力资源的方式。

方式一:终端确定激活或去激活第一算力资源。

在该方式中,终端可以自主确定激活或去激活第一算力资源。例如,终端可以基于第一预设条件,激活第一算力资源。例如,终端可以在满足第一预设条件的情况下,激活第一算力资源。其中,第一预设条件可以是预配置的、协议预定义的或者基站指示的。可选的,第一预设条件可以包括以下中的一项或多项。

条件 1: 根据业务的服务质量 (quality of service, QoS) 要求,确定激活或者去激活第一算力资源。

例如,终端可以预估业务的 QoS, 确定是否能够满足业务的 QoS 要求,从而激活或去激活第一算力资源。举例来说,终端基于第二算力资源执行第一业务,而终端确定无法达到第一业务的 QoS 的要求,那么终端可以激活第一算力资源。又例如,如果终端已激活了第一算力资源和第二算力资源,如果终端确定基于第二算力资源就能够满足第一业务的 QoS 的要求,那么为了节省能耗终端可以去激活第一算力资源。

条件 2: 根据第二算力资源的负载,确定激活或者去激活第一算力资源。

例如,终端可以比较第二算力资源的负载与第一预设阈值的大小关系,确定需要激活第一算力资源。举例来说,在第二算力资源的负载大于或大于等于第一预设阈值的情况下,终端可以确定需要激活第一算力资源。

又例如,第二算力资源的负载较大,接近最大值,那么终端可以激活第一算力资源。又例如,如果终端已激活了第一算力资源和第二算力资源,但第二算力资源的负载较低,那么为了节省能耗终端可以去激活第一算力资源。

条件 3: 基于业务信息与算力资源的关联关系,确定第一算力资源用于执行第一业务或确定第二算力资源不可用于执行第一业务。

例如,终端可以获取白名单,该白名单可以包括业务信息与算力资源的关联关系,有关白名单的描述可以参考图 3 所示的实施例,此处不再赘述。终端根据白名单,确定第一算力资源可用于执行第一业务,且第一业务具有算力需求的情况下,终端可以确定需要激活第一算力资源。

又例如,终端可以获取黑名单,该黑名单可以包括业务信息与算力资源的关联关系,有关黑名单的描述可以参考图 3 所示的实施例,此处不再赘述。终端根据黑名单,确定第二算力资源不可用于执行第一业务,且第一业务具有算力需求的情况下,终端可以确定需要激活第一算力资源。

终端在满足上述条件 1~条件 3 中至少一个条件的情况下,可以激活第一算力资源。在一个示例中,终端可以向基站告知激活的第一算力资源的信息。例如,终端可以向基站发送第一算力资源的标识。可选的,终端可以通过用于激活算力资源的信令或信元向基站发送第一算力资源的信息,也就是终端可以通过专用于激活算力资源的信令或信源向基站发送第一算力资源的信息。或者,终端可以向基站发送第一算力资源的信息以及对应的动作,如激活,也就是说终端可以通过通用的信令或信源携带第一算力资源的信息以及对应的动作向基站告知激活了第一算力资源。

又一种可能的情况中,终端可以基于第二预设条件,去激活第一算力资源。例如,终端可以在满足第二预设条件的情况下,去激活第一算力资源。其中,第二预设条件可以是预配置的、协议预定义的或者基站指示的。可选的,第二预设条件可以包括以下中的一项或多项。

条件 4: 比较第二算力资源的负载与第二预设阈值的大小关系,确定需要去激活第一算力资源。

例如,在第二算力资源的负载小于或小于等于第二预设阈值的情况下,终端可以确定需要去激活第一算力资源。又例如,在第二算力资源的负载小于或小于等于第二预设阈值的情况达到第一预设时长的情况下,终端可以确定需要去激活第一算力资源。

条件 5: 比较第二算力资源的负载与第三预设阈值的大小关系,以及比较第一算力资源与第四预设阈值的大小关系,确定需要去激活第一算力资源。

例如,在第二算力资源的负载小于或小于等于第三预设阈值,且第一算力资源的负载小于或小于等于第四预设阈值的情况下,终端可以确定需要去激活第一算力资源。又例如,在第二算力资源的负载小于或小于等于第三预设阈值,且第一算力资源的负载小于或小于等于第四预设阈值的情况达到第二预设时长的情况下,终端可以确定需要去激活第一算力资源。

需要说明的是,上述第一预设阈值、第二预设阈值、第三预设阈值和第四预设阈值可以相同也可以不同,第一预设时长和第二预设时长可以相同也可以不同,本申请不做具体限定。

条件 6: 第一算力资源的开始时长达到第二时长。

其中, 第一算力资源的开始时长可以理解成激活、开启或启动第一算力资源的时长。例如, 在激活第一算力资源时终端可以开始计时, 在计时时长达到第二时长的情况下, 终端可以认为满足上述条件 6。又例如, 终端在激活第一算力资源时可以开启定时器 (timer), 在定时器超时的情况下, 终端可以认为满足上述条件 6。

条件 7: 基于业务信息与算力资源的关联关系, 确定需要去激活第一算力资源。

例如, 终端可以获取白名单, 该白名单可以包括业务信息与算力资源的关联关系, 有关白名单的描述可以参考图 3 所示的实施例, 此处不再赘述。终端根据白名单, 确定第一算力资源用于执行第二业务, 且第二业务无算力需求的情况下, 终端可以确定需要激活第一算力资源。

又例如, 终端可以获取黑名单, 该黑名单可以包括业务信息与算力资源的关联关系, 有关黑名单的描述可以参考图 3 所示的实施例, 此处不再赘述。终端根据黑名单, 确定第一算力资源不可用于执行第一业务的情况下, 终端可以确定需要激活第一算力资源。

终端在满足上述条件 4~条件 7 中至少一个条件的情况下, 可以去激活第一算力资源。在一个示例中, 终端可以向基站告知去激活的第一算力资源的信息。例如, 终端可以基站发送第一算力资源的标识。可选的, 终端可以通过用于去激活算力资源的信令或信元向基站发送第一算力资源的信息。或者, 终端可以向基站发送第一算力资源的信息以及对应的动作, 如去激活。

基于上述方案, 终端可以自行决策或者基于条件决策激活或去激活第一算力资源, 从而可以节省终端的能耗以及减少业务的时延。

方式二: 终端根据基站的指示激活或去激活第一算力资源。

例如, 终端可以接收基站的信令信息, 该信令信息可以指示激活或去激活第一算力资源。终端可以根据该信令信息的指示, 激活或去激活第一算力资源。

上述信令信息可以是单信令, 如 RRC、MAC CE 或 DCI 中的一个。或者, 上述信令信息可以是信令组合。举例来说, 基站可以通过 RRC+MAC CE 或者 RRC+DCI 的方式指示激活或去激活第一算力。例如, 基站可以通过 RRC 向终端发送配置, 并为每个配置分配一个标识/索引 (index)。比如, 基站通过 RRC 向终端发送配置, 其中激活 CPU1 对应索引 1, 激活 CPU2 对应索引 2, 去激活/关闭 CPU1 对应索引 3, 去激活/关闭 CPU2 对应索引 4。如果基站想要让终端激活 CPU1, 则基站可以向终端发送 MAC CE 或 DCI, 该 MAC CE 或 DCI 可以携带索引 1, 从而终端可以确定激活 CPU1。

可选的, 上述单信令或信令组合中 RRC 携带的算力资源的信息可以是算力资源标识, 如 GPU ID/CPU ID 或算力资源标识列表 (或组), 如 CPU ID list/GPU ID list。举例来说, 基站通过 RRC、MAC CE 或 DCI 中的一个向终端指示激活或去激活的算力资源, 该 RRC、MAC CE 或 DCI 中的一个可以携带 CPU ID/GPU ID 或 CPU ID 表/GPU ID 表。这样, 终端可以激活或去激活 CPU ID/GPU ID 或 CPU ID 表/GPU ID 表指示的 CPU/GPU。例如, 基站通过 RRC 向终端发送配置, 每个配置可以携带一个 CPU ID/GPU ID, 或 CPU ID 表/GPU ID 表, 从而指示每个配置所针对的 CPU/GPU。

可选的, 上述单信令或信令组合中 RRC 指示激活或去激活的算力资源可以通过比特位图 (bit map) 实现。例如, UE 共有 8 个 CPU/GPU, 假设编号设为 0~7, 对应 8 个比特。其中, 当某个比特的取值为 0 时, 可以表示该 CPU/GPU 去激活, 当某个比特的取值为 1 时, 可以表示该 CPU/GPU 激活。例如, 00001000, 可以表示激活编号为 4 的 CPU/GPU, 其余 CPU/GPU 去激活。反之亦可, 当某个比特的取值为 1 时, 可以表示该 CPU/GPU 去激活, 当某个比特的取值为 0 时, 可以表示该 CPU/GPU 激活。例如, 11110111, 可以表示激活编号为 4 的 CPU/GPU, 其余 CPU/GPU 去激活。

一种可能的情况中, 终端在满足上述条件 1~条件 7 中任意一个条件的情况下, 终端可以向基站发送满足的条件, 从而基站可以根据终端满足的条件来指示终端激活或去激活哪些算力资源。例如, 终端在满足条件 1 的情况下, 如在第二算力资源的负载大于或大于等于预设阈值的情况下, 终端可以向基站发送指示信息, 指示终端满足条件 1, 如终端可以向基站发送条件 1 对应的索引。这样, 基站可以向终端指示激活第一算力资源。

另一种可能的情况中, 终端可以向基站上报事件 (event), 从而基站可以根据事件来指示终端激活或去激活哪些算力资源。例如, 基站可以为终端配置触发上报事件的条件, 如条件 1~条件 6 中的一项或多项。在终端满足触发上报事件的条件, 终端可以向基站上报事件。例如, 终端在满足条件 1 的情况下, 如在第二算力资源的负载大于或大于等于预设阈值的情况下, 终端可以向基站上报该事件,

指示终端的第二算力资源的负载大于或大于等于预设阈值,如终端可以向基站发送该事件的索引。这样,基站可以向终端指示激活第一算力资源。

可选的,终端可以向基站发送指示信息,指示或者说建议基站激活或去激活哪些算力资源。例如,终端满足上述条件 1~条件 7 中一个或多个条件的情况下,可以向基站发送指示信息,指示激活或去激活哪些算力资源。举例来说,终端在满足条件 2 的情况下,如终端确定第一算力资源用于执行第一业务的情况下,可以向基站发送激活第一算力资源的指示信息,如可以发送第一算力资源的标识和对应的动作,如激活。又例如,终端预估业务的 QoS 是否能够满足,从而向基站发送指示信息。举例来说,终端基于第二算力资源执行第一业务,而终端确定无法达到第一业务的 QoS 的要求,那么终端向基站发送激活第一算力资源的指示信息。

基于上述方案,终端可以基于基站的指示,激活或去激活算力资源,从而可以减少终端的能耗或减少业务的时延。

需要说明的是,上述图 5 所示的实施例可以作为一个单独的实施例实施,或者本领域技术人员也可以选择图 5 和图 2 所示的实施例中的步骤组合成为一个完整的实施例实施。例如,通过图 5 所示的实施例,终端可以激活第一算力资源。终端可以采用图 2 所示的实施例来为业务分配第一算力资源,如分配激活的 CPU/GPU 的时间片,本申请不做具体限定。

本文通过上述实施例介绍了在终端存在多个 CPU/GPU 的情况下,终端激活或去激活哪些 CPU/GPU 的方法。目前,在无线网络中可以共享的算力资源可以包括调制解调器(modem)集成的算力资源,如 CPU/GPU。而调制解调器无法使用终端通用的算力资源,如应用处理器(application processing unit, APU)。鉴于此,本申请实施例还提供另一种资源调度方法,用于调制解调器请求终端通用的算力资源来执行业务的计算需求。

参阅图 6,为本申请实施例提供的一种资源调度方法,可以包括以下操作。

S601: 第一装置向第二装置发送请求信息。

相应的,第二装置接收来自第一装置的请求信息。

其中,请求信息可以用于请求第二装置的算力资源。请求信息中可以包括所需的计算量、服务质量、业务信息和算力资源标识中的一项或多项。可选的,第一装置和第二装置可以应用于不同设备的装置,如第一装置可以是应用于第一设备的装置,如芯片或电路模块,或者第一装置可以是第一设备,第二装置可以是应用于第二设备的装置,如芯片或电路模块,或者第二装置可以是第二设备。

可选的,第一装置和第二装置是应用于同一设备的不同装置,如第一装置可以是终端设备中的调制解调器,第二装置可以是终端设备中的 APU。

在一种可能的实现方式中,参阅图 7,本申请实施例提供一种第一装置与第二装置之间的应用程序接口(application programming interface, API)。S601 中,第一装置可以通过 API 向第二装置发送请求信息。本文中提供了三种不同的 API,不同的 API 所需要的参数不同,以下进行介绍。

API1: 计算量(computation)。其中,第一装置通过 API1 向第二装置发送请求信息的情况下,请求信息中需要携带业务的计算量,如计算量的最大值(tops)或计算量占第二装置的算力资源的百分比。

基于 API1,第一装置可以向第二装置请求满足业务的计算量需求的算力资源。

可选的,API1 还需要业务的 QoS 和算力资源标识中的一个或多个。其中,QoS 可以包括计算时延要求或优先级、算力资源标识可以包括 CPU ID/GPU ID 或 CPU/GPU 时间片的位置。第一装置通过 API1 向第二装置发送请求信息的情况下,请求信息中需要携带业务的 QoS 和算力资源标识中的一个或多个以及业务的计算量。基于 API1,第一装置可以向第二装置请求适用于满足业务的 QoS 要求和/或符合算力资源标识的算力资源,可以提高效率。

API2: 算力资源标识,如 CPU/GPU 时间片的位置。可选的,CPU/GPU 时间片的位置可以是周期性的。第一装置通过 API2 向第二装置发送请求信息的情况下,请求信息中需要携带算力资源标识。可选的,请求信息中还可以携带周期的指示信息。基于 API2,第一装置可以向第二装置请求特定的算力资源,用于执行相应业务,特别是周期性的业务。

API3: 业务信息和算力资源标识。其中,业务信息可以包括业务标识或逻辑信道,算力资源标识可以包括 CPU ID/GPU ID 或 CPU/GPU 时间片的位置。第一装置通过 API3 向第二装置发送请求信息的情况下,请求信息中需要携带业务信息和算力资源标识。基于 API3,第一装置可以向第二装置请求适用于执行特定业务的算力资源,可以提高效率。

S602: 第二装置向第一装置发送响应信息。

相应的, 第二装置接收来自第一装置的响应信息。

其中, 响应信息可以用于指示请求算力资源成功或者请求算力资源失败。可选的, 在响应信息指示请求算力资源成功的情况下, 第一装置可以采用请求的算力资源执行业务。

5 应理解, 本申请实施例中的算力资源的标识是示例性地用于指示算力资源的方式, 还可以采用其他方式或指示信息来指示算力资源, 本申请对此不作具体限定。

需要说明的是, 图 6 所示的实施例可以作为一个单独的实施例实施, 或者本领技术人员可以选择图 2、图 5 和图 6 所示的实施例中的一些步骤组合为一个实施例实施。例如, 在图 6 所示的实施例中请求算力资源成功的情况下, 可以采用图 2 所示的实施例的方式为业务分配算力资源。又例如, 在通过图 5 10 所示的实施例激活第一算力资源的情况下, 仍无法满足业务的计算需求, 可以采用图 6 所示的实施例来请求通用的算力资源。

下面结合附图介绍本申请实施例中用来实现上述方法的通信装置。因此, 上文中的内容均可以用于后续实施例中, 重复的内容不再赘述。

图 8 为本申请实施例提供的通信装置 800 的示意性框图。该通信装置 800 可以对应实现上述各个方法实施例中由终端设备或网络设备实现的功能或者步骤。该通信装置可以包括处理单元 810。可选的, 还可以包括收发单元 820 和/或存储单元。其中, 收发单元 820 可以用于执行由终端设备或网络设备所执行的全部接收或发送操作。该存储单元可以用于存储指令(代码或者程序)和/或数据。处理单元 810 可以与该存储单元耦合, 例如, 处理单元 810 可以读取存储单元中的指令(代码或者程序)和/或数据, 15 以实现相应的方法。上述各个单元可以独立设置, 也可以部分或者全部集成。

20 在一些可能的实施方式中, 通信装置 800 能够对应实现上述方法实施例中终端设备的行为和功能。例如通信装置 800 可以为终端设备, 也可以为应用于终端设备中的部件(例如芯片或者电路)。其中, 处理单元 810 用于执行如图 2 或图 5 所示的实施例中由终端设备所执行的除了收发操作之外的全部操作。例如图 2 所示的实施例中的 S201、S202、S203, 和/或用于支持本文所描述的技术的其它过程。

例如, 处理单元 810, 用于获取算力资源。处理单元 810 还用于, 确定一个或多个业务的算力控制参数, 并基于算力控制参数为一个或多个业务分配算力资源。 25

又例如, 处理单元 810, 用于激活或去激活第一算力资源, 并基于第一算力资源和/或第二算力资源执行第一业务。其中, 第一算力资源与第二算力资源不同。

30 在一些可能的实施方式中, 通信装置 800 能够对应实现上述方法实施例中网络设备的行为和功能。例如通信装置 800 可以为网络设备, 也可以为应用于网络设备中的部件(例如芯片或者电路)。收发单元 820 可以用于执行由网络设备所执行的全部接收或发送操作。其中, 处理单元 810 用于执行由网络设备所执行的除了收发操作之外的全部操作。

例如, 处理单元 810, 用于确定用于终端设备确定算力控制参数的参数。收发单元 820, 用于向终端设备发送配置信息, 配置信息包括参数。其中, 算力控制参数用于终端设备分配算力资源。

又例如, 处理单元 810, 用于确定第一预设条件和/或第二预设条件。收发单元 820, 用于向终端设备发送第一预设条件和/或第二预设条件。第一预设条件用于终端设备确定是否激活第一算力资源, 第二预设条件用于终端设备确定是否去激活第一算力资源。有关第一预设条件和第二预设条件可以参见方法实施例的相关描述。 35

有关处理单元 810 和收发单元 820 所执行的操作, 可以参见前述方法实施例的相关描述。

40 应理解, 本申请实施例中的处理单元 810 可以由处理器或处理器相关电路组件实现, 收发单元 820 可以由收发器或收发器相关电路组件或者通信接口实现。

基于同一构思, 如图 9 所示, 本申请实施例提供一种通信装置 900。该通信装置 900 包括处理器 910。可选的, 通信装置 900 还可以包括存储器 920, 用于存储处理器 910 执行的指令或存储处理器 910 运行指令所需要的输入数据或存储处理器 910 运行指令后产生的数据。处理器 910 可以通过存储器 920 存储的指令实现上述方法实施例所示的方法。

45 基于同一构思, 如图 10 所示, 本申请实施例提供一种通信装置 1000, 该通信装置 1000 可以是芯片或者芯片系统。可选的, 在本申请实施例中芯片系统可以由芯片构成, 也可以包含芯片和其他分立器件。

通信装置 1000 可以包括至少一个处理器 1010, 该处理器 1010 与存储器耦合, 可选的, 存储器可

以位于该装置之内，也可以位于该装置之外。例如，通信装置 1000 还可以包括至少一个存储器 1020。存储器 1020 保存实施上述任一实施例中的必要计算机程序、配置信息、计算机程序或指令和/或数据；处理器 1010 可能执行存储器 1020 中存储的计算机程序，完成上述任一实施例中的方法。

本申请实施例中的耦合是装置、单元或模块之间的间接耦合或通信连接，可以是电性，机械或其它的形式，用于装置、单元或模块之间的信息交互。处理器 1010 可能和存储器 1020 协同操作。本申请实施例中不限定上述收发器 1030、处理器 1010 以及存储器 1020 之间的具体连接介质。

通信装置 1000 中还可以包括收发器 1030，通信装置 1000 可以通过收发器 1030 和其它设备进行信息交互。收发器 1030 可以是电路、总线、收发器或者其它任意可以用于进行信息交互的装置，或称为信号收发单元。如图 10 所示，该收发器 1030 包括发射机 1031、接收机 1032 和天线 1033。此外，当该通信装置 1000 为芯片类的装置或者电路时，该通信装置 1000 中的收发器也可以是输入输出电路和/或通信接口，可以输入数据（或称，接收数据）和输出数据（或称，发送数据），处理器为集成的处理器或者微处理器或者集成电路，处理器可以根据输入数据确定输出数据。

在一种可能的实施方式中，该通信装置 1000 可以应用于终端设备，具体通信装置 1000 可以是终端设备，也可以是能够支持终端设备实现上述涉及的任一实施例中的终端设备的功能的装置。存储器 1020 保存实现上述任一实施例中的终端设备的功能的必要计算机程序、计算机程序或指令和/或数据。处理器 1010 可执行存储器 1020 存储的计算机程序，完成上述任一实施例中终端设备执行的方法。

在另一种可能的实施方式中，该通信装置 1000 可以应用于网络设备，具体通信装置 1000 可以是终端设备，也可以是能够支持网络设备，实现上述涉及的任一实施例中的网络设备的功能的装置。存储器 1020 保存实现上述任一实施例中的网络设备的功能的必要计算机程序、计算机程序或指令和/或数据。处理器 1010 可执行存储器 1020 存储的计算机程序，完成上述任一实施例中网络设备执行的方法。

由于本实施例提供的通信装置 1000 可应用于终端设备，完成上述终端设备执行的方法，或者应用于网络设备，完成网络设备执行的方法。因此其所能获得的技术效果可参考上述方法实施例，在此不再赘述。

在本申请实施例中，处理器可以是通用处理器、数字信号处理器、专用集成电路、现场可编程门阵列或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件，可以实施或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件处理器执行完成，或者用处理器中的硬件及软件模块组合执行完成。

在本申请实施例中，存储器可以是非易失性存储器，比如硬盘（hard disk drive，HDD）或固态硬盘（solid-state drive，SSD）等，还可以是易失性存储器（volatile memory），例如随机存取存储器（random-access memory，RAM）。存储器还可以是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质，但不限于此。本申请实施例中的存储器还可以是电路或者其它任意能够实施存储功能的装置，用于存储计算机程序、计算机程序或指令和/或数据。

基于以上实施例，参见图 11，本申请实施例还提供另一种通信装置 1100，包括：输入输出接口 1110 和逻辑电路 1120；输入输出接口 1110，用于接收代码指令并传输至逻辑电路 1120；逻辑电路 1120，用于运行代码指令以执行上述任一实施例中终端设备或者网络设备执行的方法。

由于本实施例提供的通信装置 1100 可应用于终端设备或网络设备，执行上述终端设备或网络设备所执行的方法，或者应用于终端设备或网络设备，完成终端设备或网络设备执行的方法。因此其所能获得的技术效果可参考上述方法实施例，在此不再赘述。

基于以上实施例，本申请实施例还提供一种通信系统。该通信系统包括至少一个应用于终端设备的通信装置和至少一个应用于网络设备的通信装置。所能获得的技术效果可参考上述方法实施例，在此不再赘述。

基于以上实施例，本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序或指令，当指令被执行时，使上述任一实施例中终端设备执行的方法被实施或者网络设备执行的方法被实施。该计算机可读存储介质可以包括：U 盘、移动硬盘、只读存储器、随机存取存储器、磁盘或者光盘等各种可以存储程序代码的介质。

为了实现上述图 8~图 11 的通信装置的功能，本申请实施例还提供一种芯片，包括处理器，用于支持该通信装置实现上述方法实施例中终端设备或者网络设备所涉及的功能。在一种可能的设计中，该芯

片与存储器连接或者该芯片包括存储器,该存储器用于保存该通信装置必要的计算机程序或指令和数据。

本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序或指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序或指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序或指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序或指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

显然,本领域的技术人员可以对本申请实施例进行各种改动和变型而不脱离本申请实施例的范围。这样,倘若本申请实施例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

1.一种资源调度方法,其特征在于,包括:

获取算力资源;

确定一个或多个业务的算力控制参数;

5 基于所述算力控制参数为所述一个或多个业务分配所述算力资源。

2.根据权利要求1所述的方法,其特征在于,还包括:

接收来自网络设备的配置信息,所述配置信息用于确定所述算力控制参数。

3.根据权利要求2所述的方法,其特征在于,所述配置信息包括保证时间窗和第一速率,所述算力
10 控制参数是所述第一速率和所述保证时间窗的乘积,所述第一速率是单位时间内分配的算力大小PCR,
所述保证时间窗用于确定所述算力控制参数的时长。

4.根据权利要求1所述的方法,其特征在于,所述算力控制参数是所述一个或多个业务的第二速率
和所述一个或多个业务的第一时长的乘积;

其中,所述第二速率是单位时间内分配的算力大小,所述第一时长是本次确定所述算力控制参数
的时间与上一次确定所述算力控制参数的时间之间的时长。

5.根据权利要求3和4所述的方法,其特征在于,在所述第二速率和所述第一时长的乘积大于或等
15 于所述第一速率和保证时间窗的乘积的情况下,所述算力控制参数为所述第一速率和保证时间窗的乘积;

在所述第二速率和所述第一时长的乘积小于所述第一速率和保证时间窗的乘积的情况下,所述算
力控制参数为所述第二速率和所述第一时长的乘积。

6.根据权利要求1~5任一所述的方法,其特征在于,所述基于所述算力控制参数为一个或多个业
20 务分配算力资源,包括:

基于所述算力控制参数和所述一个或多个业务的优先级,为所述一个或多个业务分配算力资源。

7.根据权利要求6所述的方法,其特征在于,所述基于所述算力控制参数为一个或多个业务分配算
力资源,包括:

为所述一个或多个业务中优先级最高,且所述算力控制参数大于0的业务分配算力资源。

8.根据权利要求6所述的方法,其特征在于,所述基于所述算力控制参数为一个或多个业务分配算
25 力资源,包括:

在所述一个或多个业务中存在速率无穷大的业务的情况下,为所述速率无穷大的业务分配算力资
源。

9.根据权利要求6~8任一所述的方法,其特征在于,还包括:

30 在存在剩余的算力资源的情况下,丢弃所述剩余的算力资源。

10.根据权利要求9所述的方法,其特征在于,所述丢弃所述剩余的算力资源,包括:

向网络设备发送所述剩余的算力资源的指示信息,所述剩余的算力资源的指示信息用于指示请求
去激活所述剩余的算力资源。

11.根据权利要求6~8任一所述的方法,其特征在于,还包括:

35 在存在剩余的算力资源的情况下,为所述一个或多个业务中算力控制参数大于0的业务,分配所
述剩余的算力资源。

12.根据权利要求11所述的方法,其特征在于,所述在存在剩余的算力资源的情况下,为所述一个
或多个业务中算力控制参数大于0的业务,分配所述剩余的算力资源,包括:

40 在存在剩余的算力资源的情况下,按照所述一个或多个业务中算力控制参数大于0的业务的优先
级由高到低的顺序,分配所述剩余的算力资源。

13.根据权利要求11或12所述的方法,其特征在于,还包括:

在存在剩余的算力资源的情况下,且所述一个或多个业务请求的算力资源大小为0的情况下,为
所述一个或多个业务中所述算力控制参数小于0的业务分配所述剩余的算力资源。

14.根据权利要求1~13任一所述的方法,其特征在于,还包括:

45 更新所述算力控制参数,更新的算力控制参数是基于所述算力控制参数以及为对应业务分配的算
力资源确定的。

15.根据权利要求1~14任一所述的方法,其特征在于,还包括:

接收业务与算力资源的关联关系；所述关联关系用于指示所述算力资源用于执行所述业务，或者所述关联关系用于指示所述算力资源不可用于执行所述业务；

所述基于所述算力控制参数为所述一个或多个业务分配算力资源，包括：

基于所述关联关系和/或所述算力控制参数，为所述一个或多个业务分配算力资源。

5 16.一种资源调度方法，其特征在于，包括：

确定配置信息，所述配置信息用于终端设备确定算力控制参数；

向所述终端设备发送所述配置信息；

其中，所述算力控制参数用于所述终端设备分配算力资源。

10 17.根据权利要求 16 所述的方法，其特征在于，所述配置信息包括保证时间窗和速率，所述算力控制参数是所述速率和所述保证时间窗的乘积，所述速率是单位时间内分配的算力大小 PCR，所述保证时间窗用于确定所述算力控制参数的时长。

18.根据权利要求 16 或 17 所述的方法，其特征在于，还包括：

向所述终端设备发送业务与算力资源的关联关系；所述关联关系用于指示所述算力资源用于执行所述业务，或者所述关联关系用于指示所述算力资源不可用于执行所述业务。

15 19.一种通信装置，其特征在于，包括用于执行如权利要求 1~15 中任一所述的方法的单元，或包括用于执行如权利要求 16~18 中任一所述的方法的单元。

20.一种通信装置，其特征在于，包括：处理器和存储器；

所述存储器，用于存储计算机程序或指令；

20 所述处理器，用于执行存储器中的计算机程序或指令，使所述装置执行如权利要求 1~15 中任一项所述的方法，或使所述装置执行如权利要求 16~18 中任一项所述的方法。

21.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令在被电子装置调用时，使所述电子装置执行如权利要求 1~15 中任一项所述的方法，或者使所述电子装置执行如权利要求 16~18 中任一项所述的方法。

25 22.一种通信系统，其特征在于，包括用于执行如权利要求 1~15 中任一项所述的方法的装置和用于执行如权利要求 16~18 中任一项所述的方法的装置。

23.一种芯片系统，其特征在于，所述芯片系统包括：

通信接口；

30 处理器，用于通过所述通信接口调用并运行所述指令，使得安装有所述芯片系统的设备执行如权利要求 1~15 中任一项所述的方法，或使得安装有所述芯片系统的设备执行如权利要求 16~18 中任一项所述的方法。

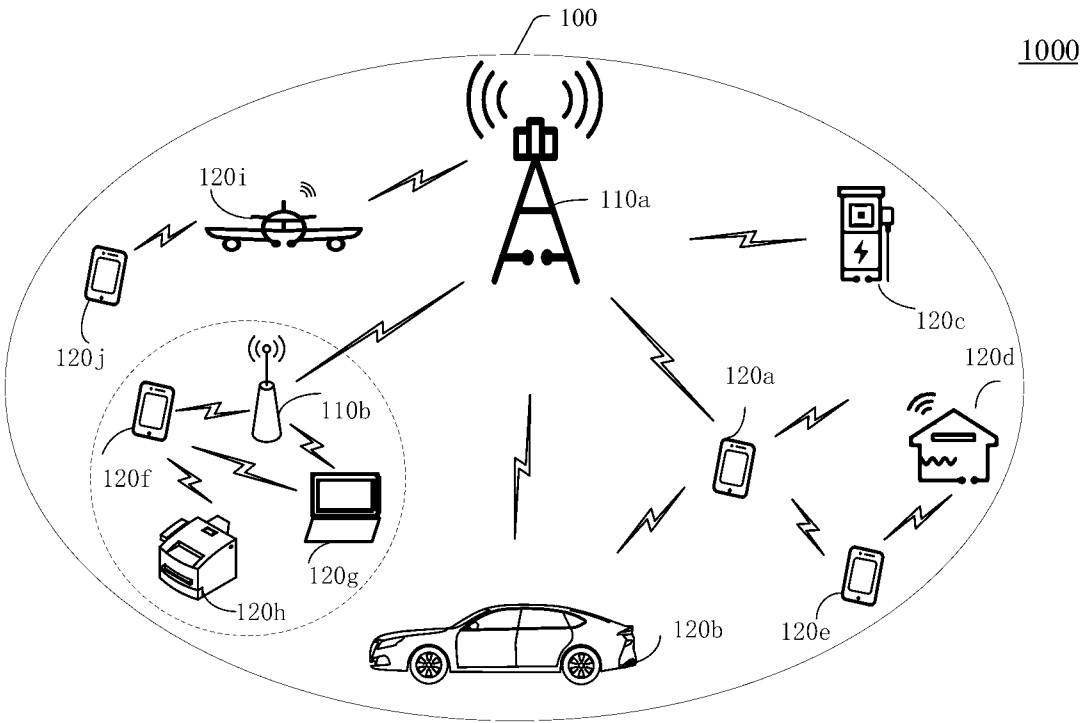


图 1

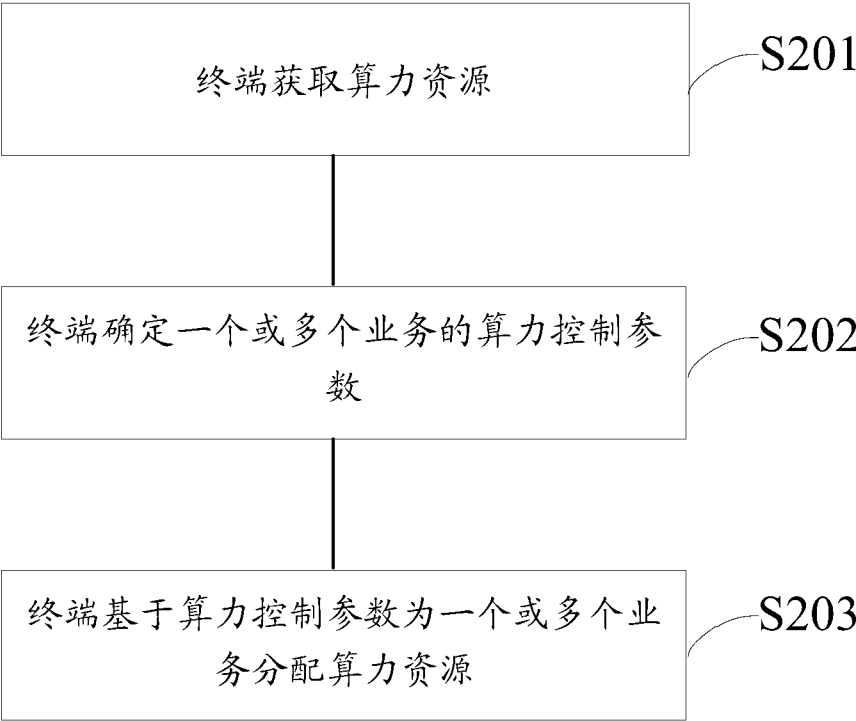


图 2

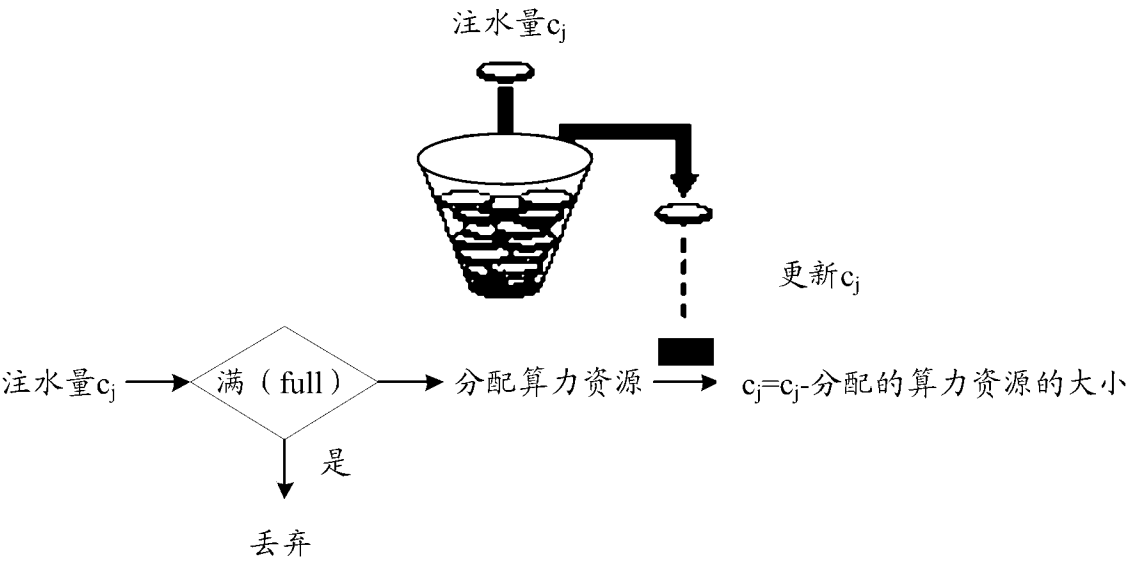


图 3

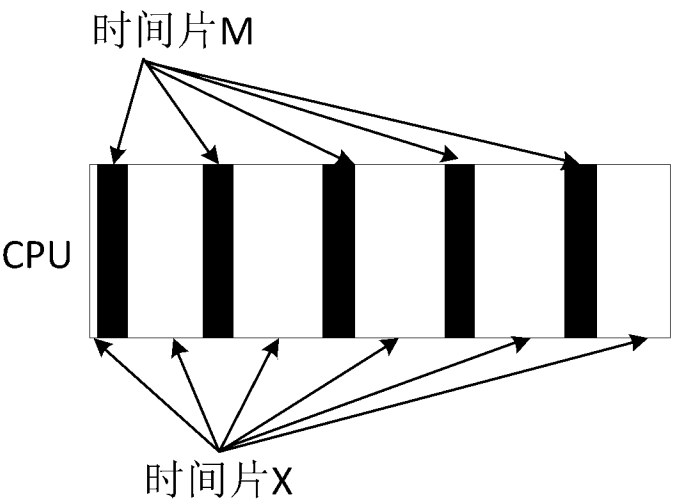


图 4

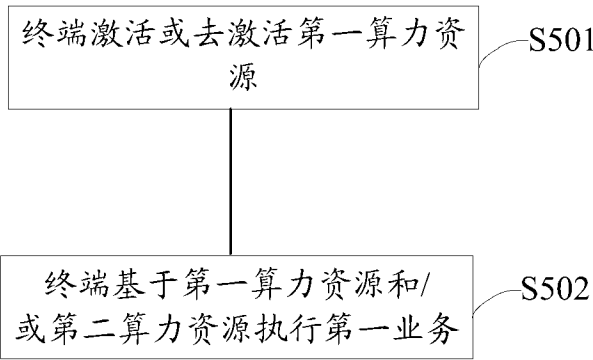


图 5

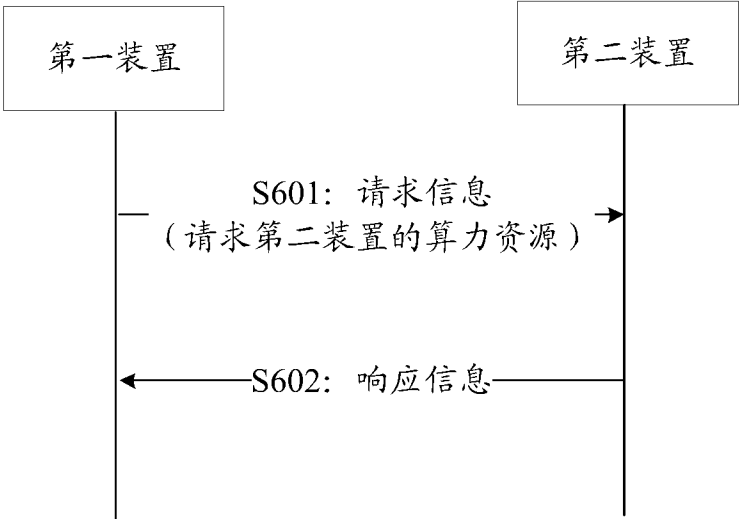


图 6

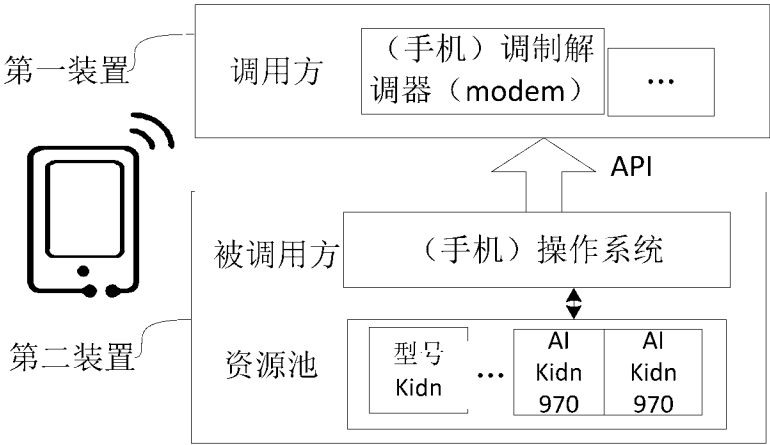


图 7

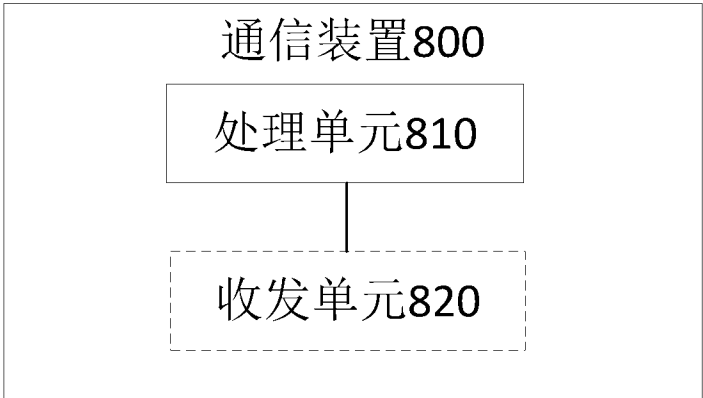


图 8

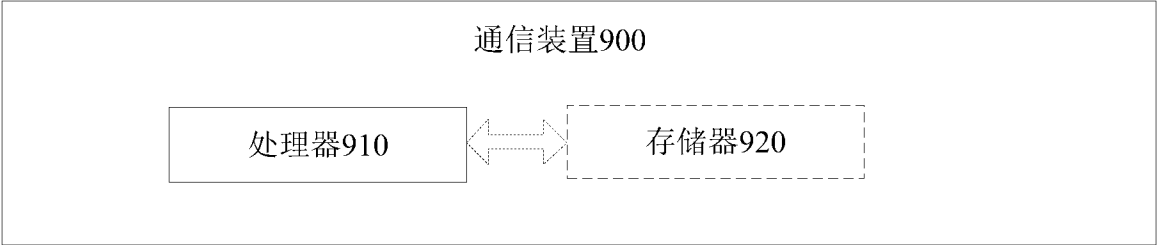


图 9

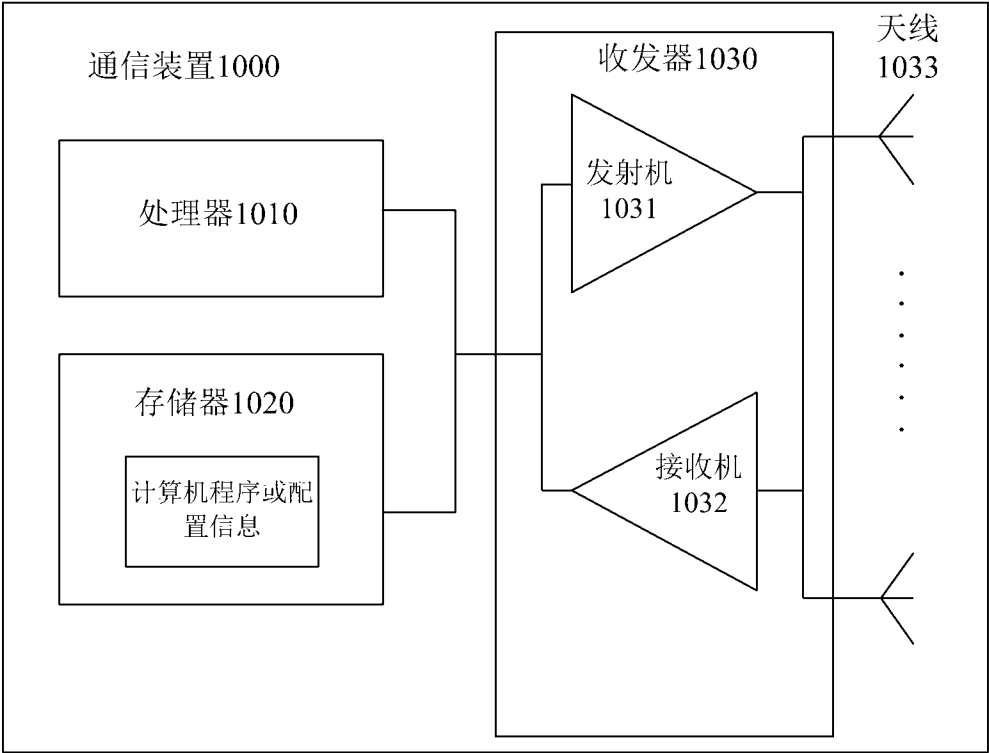


图 10

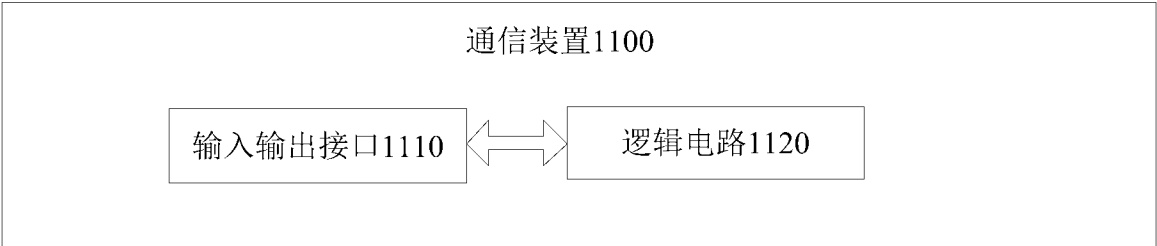


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/112968

A. CLASSIFICATION OF SUBJECT MATTER

H04L67/10(2022.01)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)

IPC: H04L H04W

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)

CNTEXT, ENTXT, ENTXTC, WPABS, DWPI, VCN, CNKI, IEEE: 算力, 资源, 分配, 配置, 保证时间窗, bsd, 桶尺寸时间区间, Bucket Size Duration, 令牌, 速率, 时长, 占比, 乘积, 剩余, 去激活, 丢, 优先级, 业务, 任务, task, computing, resource, allocation, configuration, product, rate, ratio, remain, deactivate, drop, priority

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111800877 A (CHINA UNITED NETWORK COMMUNICATIONS GROUP CO., LTD.) 20 October 2020 (2020-10-20) description, paragraphs 2-36	1-2, 6, 9-10, 14, 16, 19-23
Y	CN 111800877 A (CHINA UNITED NETWORK COMMUNICATIONS GROUP CO., LTD.) 20 October 2020 (2020-10-20) description, paragraphs 2-36	3-5, 7-8, 11- 13, 15, 17-23
Y	CN 111669831 A (CHINA UNITED NETWORK COMMUNICATIONS GROUP CO., LTD.) 15 September 2020 (2020-09-15) description, paragraphs 60-126	3-5, 7-8, 11- 13, 15, 17-23
A	CN 116302477 A (JINAN INSPUR DATA TECHNOLOGY CO., LTD.) 23 June 2023 (2023-06-23) entire document	1-23
A	CN 115484620 A (HUAWEI TECHNOLOGIES CO., LTD.) 16 December 2022 (2022-12-16) entire document	1-23



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

25 April 2024

Date of mailing of the international search report

06 May 2024

Name and mailing address of the ISA/CN

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

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/112968

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2023239868 A1 (NOKIA TECHNOLOGIES OY) 27 July 2023 (2023-07-27) entire document	1-23
A	WO 2019069464 A1 (NTT DOCOMO, INC.) 11 April 2019 (2019-04-11) entire document	1-23

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/112968

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111800877	A	20 October 2020	None			
CN	111669831	A	15 September 2020	None			
CN	116302477	A	23 June 2023	None			
CN	115484620	A	16 December 2022	WO	2022252859	A1	08 December 2022
				EP	4333533	A1	06 March 2024
				US	2024106764	A1	28 March 2024
US	2023239868	A1	27 July 2023	WO	2022029524	A1	10 February 2022
WO	2019069464	A1	11 April 2019	CA	3078521	A1	06 April 2019
				JPWO	2019069464	A1	19 November 2020
				JP	7121031	B2	17 August 2022
				US	2020296713	A1	17 September 2020
				US	11696288	B2	04 July 2023
				BR	112020006771	A2	06 October 2020
				EP	3694270	A1	12 August 2020
				EP	3694270	A4	12 May 2021

A. 主题的分类

H04L67/10(2022.01)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)

IPC: H04L H04W

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))

CNTEXT, ENTXT, ENTXTC, WPABS, DWPI, VCN, CNKI, IEEE: 算力, 资源, 分配, 配置, 保证时间窗, bsd, 桶尺寸时间区间, Bucket Size Duration, 令牌, 速率, 时长, 占比, 乘积, 剩余, 去激活, 丢, 优先级, 业务, 任务, task, computing, resource, allocation, configuration, product, rate, ratio, remain, deactivate, drop, priority

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 111800877 A (中国联合网络通信集团有限公司) 2020年10月20日 (2020 - 10 - 20) 说明书第2-36段	1-2、6、9-10、14、16、19-23
Y	CN 111800877 A (中国联合网络通信集团有限公司) 2020年10月20日 (2020 - 10 - 20) 说明书第2-36段	3-5、7-8、11-13、15、17-23
Y	CN 111669831 A (中国联合网络通信集团有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第60-126段	3-5、7-8、11-13、15、17-23
A	CN 116302477 A (济南浪潮数据技术有限公司) 2023年6月23日 (2023 - 06 - 23) 全文	1-23
A	CN 115484620 A (华为技术有限公司) 2022年12月16日 (2022 - 12 - 16) 全文	1-23
A	US 2023239868 A1 (NOKIA TECHNOLOGIES OY) 2023年7月27日 (2023 - 07 - 27) 全文	1-23
A	WO 2019069464 A1 (NTT DOCOMO, INC.) 2019年4月11日 (2019 - 04 - 11) 全文	1-23

☐ 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

“D” 申请人在国际申请中引证的文件

“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

2024年4月25日

国际检索报告邮寄日期

2024年5月6日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

张德珍

电话号码 (+86) 010-53961658

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/112968

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	111800877	A	2020年10月20日	无			
CN	111669831	A	2020年9月15日	无			
CN	116302477	A	2023年6月23日	无			
CN	115484620	A	2022年12月16日	WO	2022252859	A1	2022年12月8日
				EP	4333533	A1	2024年3月6日
				US	2024106764	A1	2024年3月28日
US	2023239868	A1	2023年7月27日	WO	2022029524	A1	2022年2月10日
WO	2019069464	A1	2019年4月11日	CA	3078521	A1	2019年4月6日
				JPWO	2019069464	A1	2020年11月19日
				JP	7121031	B2	2022年8月17日
				US	2020296713	A1	2020年9月17日
				US	11696288	B2	2023年7月4日
				BR	112020006771	A2	2020年10月6日
				EP	3694270	A1	2020年8月12日
				EP	3694270	A4	2021年5月12日