中国云体系产业创新战略联盟
China Cloud System Pioneer Strategic Alliance

云计算战略联盟技术标准

HB/P-2020-0003

大规模网络信息服务平台的在线任务动态调度框架及流程技术标准

The Technology Standards of Dynamic Scheduling Framework and Workflow for Online Tasks in Information Service Platform of Large Scale Network

编制单位：同济大学、东华大学

发布时间：2020-09-01

前　言

《大规模网络信息服务平台在线任务动态调度框架及流程技术标准》由以下5部分构成：
——第1部分：范围；
——第2部分：规范性引用文件；
——第3部分：术语和定义；
——第4部分：动态在线任务调度系统的框架；
——第5部分：动态在线任务调度系统的工作流程。

本标准按照GB/T 1.1-2009给出的规则起草。
本标准由同济大学提出。
本标准由信息技术标准化技术委员会（SAC/TC180）归口。
本标准负责起草单位：同济大学
本标准参加起草单位：东华大学
本标准主要起草人：　蒋昌俊、章昭辉、丁志军、喻剑、闫春钢、张亚英

引　言

大规模网络信息服务平台，如淘宝平台、12306订票平台等，具有任务规模大任务计算时间短的特征。同时这类平台具有众多的分布式计算结点的计算平台特征。因此，大规模网络信息服务平台对任务调度系统具有高吞吐率的性能要求。

任务调度是将任务合理地分配到资源上去，达到负载均衡和高吞吐率。为了合理、充分的使用资源，必须对任务进行优化调度。目前，较为著名的任务调度器有便携式批量处理系统（PBS）、负载共享器（LSF）、负载平均器（Load Leveler）、负载均衡器（Load Balancer）、批量调度器（Batch）等。这些任务调度系统主要应用于“高性能计算”的机群系统，适合低吞吐率的大型任务计算。在大规模网络信息服务系统中，对于规模庞大的小任务而言，缺乏合适在线任务动态调度系统及其统一标准规范。为此，特制定大规模网络信息服务平台在线任务动态调度框架及流程技术标准。

大规模网络信息服务平台的在线任务动态调度框架及流程技术标准

1 范围

本标准规范了大规模网络信息服务平台的高吞吐率小任务动态调度的框架及工作流程，对其进行统一的名称规范和定义说明，并为大规模网络信息服务平台其他各项标准的编制提供参照。

本标准适用于所有大规模网络信息服务平台相关组织及其设计、研制、发行、管理、维护的产品、系统等，为行业的信息服务平台提供参照性规范。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1-2009 标准化工作导则

ISO/IEC 23006-4-2013 信息技术 多媒体服务平台技术 第4部分:基本服务

ISO/IEC 23006-1-2013 信息技术 多媒体服务平台技术 第1部分:架构

GA/T 739.1-2007 公安请求服务平台应用规范 第1部分:应用服务描述

ISO/IEC 23006-5-2013 信息技术 多媒体服务平台技术 第5部分:服务聚合

GB/T 25470-2010 制造业信息化共性技术资源服务平台功能规范

GA/T 739.2-2007 公安请求服务平台应用规范 第2部分:请求服务应用接口

GA/T 1038.3-2012 消防公共服务平台技术规范 第3部分:信息交换接口

GB/T 30290.1-2013 卫星定位车辆信息服务系统 第1部分:功能描述

GB/T 29746-2013 实时交通信息服务数据结构

AS 3965-1991 信息技术 开放系统和互联 公共管理信息服务定义

GB/T 29841.4-2013 卫星定位个人位置信息服务系统 第4部分:终端通用规范

3 术语和定义

在线任务动态调度：即接收到任务后根据当前资源状况立即进行调度。任务调度是将任务合理地分配到资源上去，达到负载均衡和高吞吐率。任务调度遵循以下两个分配原则：(1)负载均衡原则。任务应该分配到完成时间较早的处理机上，使得各处理机运行时间相当。(2)高吞吐率原则。任务应该分配到执行时间较少的处理机上。

4 动态在线任务调度系统的框架

4.1 概述

在大规模网络信息服务平台中，任务调度系统要能调度大规模的小任务同时还要满足系统高吞吐率的调度性能。其特点表现为以下几点：(1) 在大规模网络信息服务的基础平台上（如云计算平台），计算任务均是使用资源规范语言（Resource Specification Language，RSL）描述。任务调度系统先解析基础平台的RSL，然后调度。(2)调度系统与基础平台的资源分配管理（Resource Allocation Management,RAM）、元计算目录服务（Metacomputing Directory Service,MDS）相结合，充分利用基础平台提供的任务管理功能和平台资源信息。(3)调度系统面向小任务（计算时间约为几秒或几十秒），适合高任务到达频率。

4.2 任务调度系统的架构

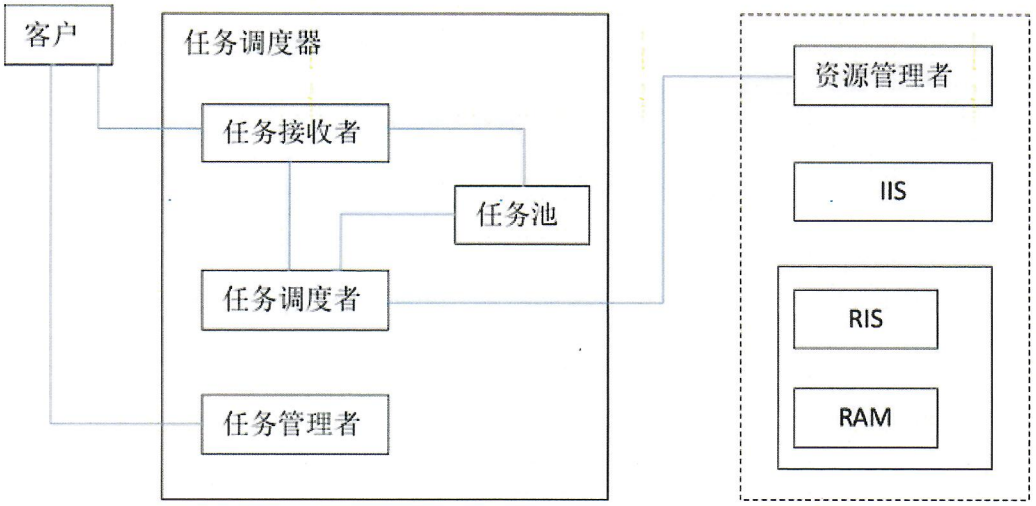

图1 任务调度系统架构图

任务调度系统的架构如图1所示。

任务调度器由任务池，任务接收者，任务调度者，任务管理者组成。任务接收者将收到的任务放入任务池中。资源管理者采集资源信息，供任务调度者使用。任务调度者负责采用调度策略调度任务池中的任务，并为每个任务创建一个任务管理者。任务管理者相当于客户代理，根据客户的需求管理任务。资源分配管理（Resource Allocation Management,RAM）模块、资源信息服务(Resource Information Service，RIS)模块、信息目录服务(Information Index Service，IIS）模块是基础平台本身提供的模块，用于采集基础平台的资源信息。

（1）任务池

如图2所示，任务池是任务接收者与任务调度者共享的一块内存，用于暂时存放任务。当任务的提交速度高于任务的调度速度时，任务在池内排队，等待调度。共享内存使用系统调用申请。任务池是一个结构体数组，每个结构体存放一个任务。为便于使用任务池，用两个链表管理对其进行管理。两个链表分别是任务链表（job_list）和空闲链表（idle_list）。任务链表中的结构体是存放作业的；空闲链表中的结构体是空的。在基础平台中，任务使用RSL描述。RSL中指明了该任务所需资源，可执行程序的全路径名及参数，输入文件、输出文件的全路径名等。客户的一个作业包括三个部分：客户机IP、端口以及RSL。IP与端口用于确定客户程序，以便将结果发送回去。RSL阐明了客户所要求进行的计算。任务池作为临界资源，为避免竞争条件（Race Condition），受信号量保护。任务接收者和任务调度者必须在临界区内访问任务池，进入或退出临界区需要分别做P（请求资源）或V（释放资源）操作。使用系统调用申请获得信号量。

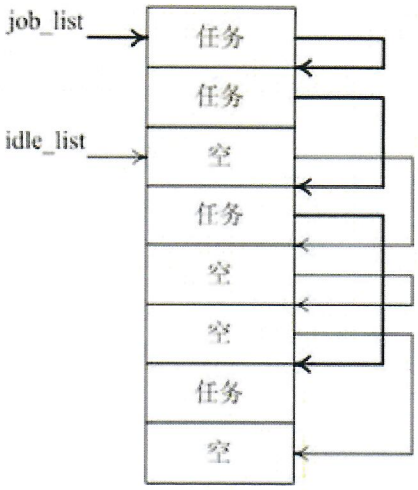

图2 调度器的任务池结构

（2）任务接收者

任务接收者是后台端口监控进程，负责接收客户提交的任务。客户使用套接字（Socket）实现与任务接收者的通信。接收者在端口收到一个任务后，进入临界区，从任务池的idle_list摘得一个空闲结点，将任务填入结点中，将结点链入job_list，退出临界区，继续监控端口。在具体实现时，任务接收者需要根据任务池的不同情况，作不同的处理。有两种特殊情况需要注意：(1)任务池已满，任务接收者无法获得空闲结点。任务接收者使用系统调用，进入挂起状态。(2)任务池已空，任务接收者将刚收到的任务存入任务池后，使用系统调用向任务调度者发送信号，将其唤醒。

任务向远程计算结点 p_k 的提交：任务调度者在提交任务之前，需考虑任务请求的资源，分以下3种情况：(1)若任务 j_i 请求的资源过多，任何一个计算节点都无法满足其需求，则调度者将 j_i 删除，并告知其提交者。(2)若 j_i 请求的资源多于计算节点 p_k 的资源，则调度者将该任务退还给任务池，调度者选择下一个任务。(3)若 p_k 的资源可以满足 j_i 对资源的需求，则可以实现匹配。在具体实现时，有两种特殊情况需要注意：(1)任务池中的任务被全部调度后，任务池为空。此时调度者使用系统调用，进入挂起状态，等待任务接收者发信号将其唤醒。(2)原本任务池已满，任务调度者调度完一个任务后，任务池中有一个空闲结点。此时任务调度者必须使用系统调用向任务接收者发送信号，将其唤醒。

(3)任务调度者

针对大规模信息服务平台的高响应率、小任务特点，调度系统应采用动态在线调度方式。当任务池中有待调度任务时：(1)任务调度者按照一定的优先级取出一个任务,解析其RSL；(2)通过资源管理者获取资源；(3)提交任务，并创建与该任务对应的任务管理者。

（4）任务管理者

基础平台提供的应用程序接口（API）可以:(1)获得任务的状态；(2)取消正在执行的任务，杀死进程，回收该进程所占用的资源。另外，调度系统应实现以下3个API，用于结果数据的返回:(1)从结果文件读取数据；(2)使用Socket发送数据；(3)打开Socket端口，接收数据。

（5）信息服务平台的任务

任务由3部分组成：客户端IP地址、客户端端口号、RSL。客户端IP地址、客户端端口号用于定位客户端程序。RSL用于描述客户请求的任务，指明任务所在的路径、参数、输入输出文件等。

5 动态在线任务调度系统的工作流程

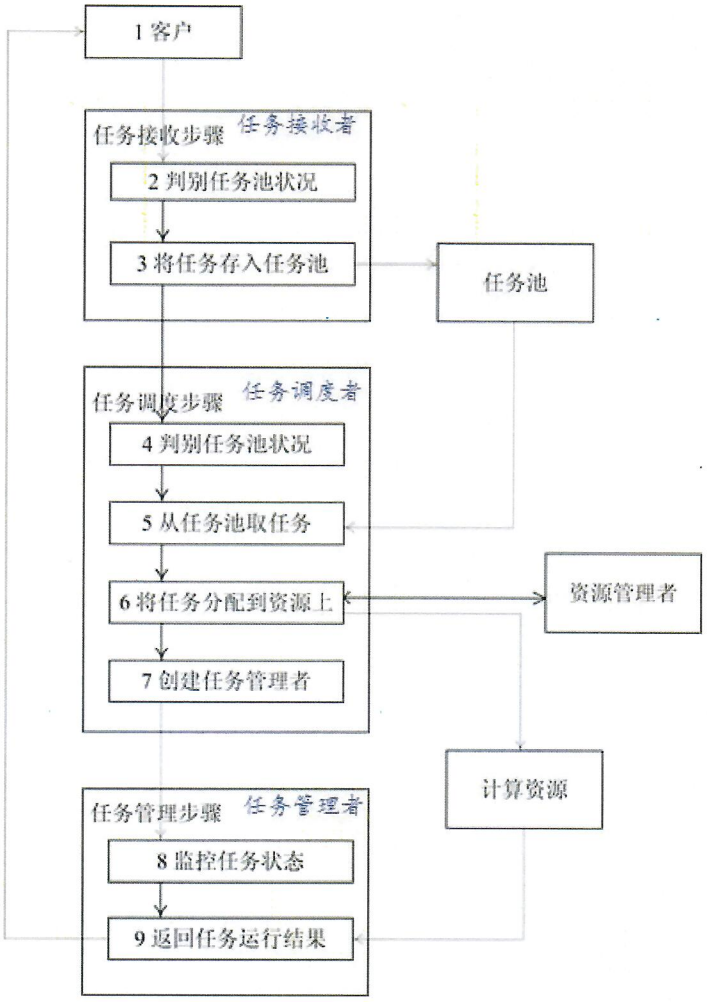

图 3 任务调度系统的工作流程

任务调度系统的工作流程如图3所示，具体流程如下：

(1)客户通过网络使用套接字Socket通信，将任务发送给任务接收者。

(2)任务接收步骤：判断任务池是否还有空间，作出相应操作，然后继续侦听端口。

a)若任务池为空或有空闲结点，任务接收者将任务放入任务池中，并发送信号给任务调度者。

b)若任务池已满，任务接收者执行系统调用，进入挂起状态。

(3)任务接收步骤：将任务加入到任务池中的job_list的队尾。

(4)任务调度步骤：判别任务池的job_list中是否有任务。

(5)任务调度步骤：若有待调度任务，则任务调度者从job_list的队首取出一个任务。

(6)任务调度步骤：解析任务的RSL语句，根据任务对资源的需求，通过资源管理者为其分配计算资源，并将任务分配到资源上。

(7) 任务调度步骤：创建子进程（任务管理者），并将刚才调度的任务的句柄交给该任务的管理者。

(8)任务管理步骤：监控任务当前状态，并且侦听端口等待客户请求。

(9)任务管理步骤：从任务的输出文件中读入结果，将结果返回给客户。

参 考 文 献

GB/T 1.1-2009 标准化工作导则

JR/T 0096.6-2012 中国金融移动支付 联网联合 第6部分：安全规范

GB 4943.1-2011 信息技术设备 安全 第1部分：通用要求

JR/T 0097-2012 中国金融移动支付 可信服务管理技术规范

ISO/IEC 23006-4-2013 信息技术 多媒体服务平台技术 第4部分:基本服务

ISO/IEC 23006-1-2013 信息技术 多媒体服务平台技术 第1部分:架构

GA/T 739.1-2007 公安请求服务平台应用规范 第1部分:应用服务描述

ISO/IEC 23006-5-2013 信息技术 多媒体服务平台技术 第5部分:服务聚合

GB/T 25470-2010 制造业信息化共性技术资源服务平台功能规范

GA/T 739.2-2007 公安请求服务平台应用规范 第2部分:请求服务应用接口

GA/T 1038.3-2012 消防公共服务平台技术规范 第3部分:信息交换接口

GB/T 30290.2-2013 卫星定位车辆信息服务系统 第2部分:车载终端与服务中心信息交换协议

GB/T 30290.1-2013 卫星定位车辆信息服务系统 第1部分:功能描述

GB/T 29746-2013 实时交通信息服务数据结构

AS 3965-1991 信息技术 开放系统和互联 公共管理信息服务定义

GB/T 29841.4-2013 卫星定位个人位置信息服务系统 第4部分:终端通用规范