基于模拟退火的多核多用户任务卸载调度

鲁 伟,宋荣方
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:针对传统的集中式网络架构存在链路负载过重、时延较长的问题,将服务器下放至靠近用户端的移动边缘计算概念孕育而生。在移动边缘计算系统中,任务卸载调度策略的好坏影响到系统时延和用户体验,因此任务卸载调度问题依旧是移动边缘计算领域中的研究热点。在移动边缘计算的多用户多核系统中,该文对用户的多个独立任务的调度策略与功率分配进行了研究。为了降低任务卸载时延,首先利用混合流水车间模型对任务卸载调度策略进行了建模,获得了系统时延的表达式,其次利用模拟退火算法对系统时延与能耗的加权和最小化的问题进行了求解,获得了最优的任务卸载甘特图。与随机任务卸载调度策略相比,所提的卸载策略可以有效降低系统时延。最后通过权重的变化,找到一个合适的权重,在不增加时延的情况下,实现了能耗的节约。

关键词:移动边缘计算;功率分配;任务卸载调度策略;混合流水车间;模拟退火算法
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2021)06-0076-05
doi:10.3969/j.issn.1673-629X.2021.06.014

Multi-core Multi-user Task Offloading Scheduling Based on Simulated Annealing Algorithm

LU Wei, SONG Rong-fang
(School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In view of the problem of excessive link load and long time delay in the traditional centralized network architecture, the concept of mobile edge computing that the server is decentralized close to the user side was born. In mobile edge computing system, the task offload scheduling strategy has a deep influence on the system delay and user experience. Therefore, task offload scheduling is still a hot topic in the field of mobile edge computing. In the multi-user and multi-core system of mobile edge computing, we study the scheduling strategy and power allocation of multiple independent tasks of users. In order to reduce the task offload delay, firstly, the hybrid flow-shop scheduling is used to model the task offload scheduling strategy and obtain the expression of the system delay. Secondly, the simulated annealing algorithm is used to minimize the weighted sum of the system delay and energy consumption. The optimal task offload Gantt chart was obtained. Compared with the random task offload scheduling strategy, the offload strategy proposed can effectively reduce the system delay. Finally, through the change of the weight, a suitable weight is found, and the energy consumption is saved without increasing the delay.

Key words: mobile edge computing; power allocation; task offload scheduling strategy; hybrid flow-shop scheduling; simulated annealing algorithm

0 引 言

随着物联网、移动互联网、大数据技术的快速发展,人类进入了一个万物互联的智能时代,移动智能终端随时随地在线,服务于移动终端上的交互式游戏、智慧城市等计算密集型的业务也正在兴起[1],这些业务需要大量的计算资源才能满足自身对低时延的要求[2]。由于移动智能设备处理能力、存储容量有限,因此大量的计算需要在云端进行[3],而云端存在较大的传输时延,当云端资源不足时,甚至存在较大的排队等待时延,这些时延严重影响了众多业务的服务质量。为了使用户能获得良好的体验,减轻云端服务器的负担,移动边缘计算(mobile edge computing,MEC)概念孕育而生[4-5]。与传统的集中式网络架构不同,MEC将边缘服务器部署在靠近用户的一端,缩短了用户与

收稿日期:2020-07-10 **修回日期**:2020-11-13
基金项目:江苏省自然科学基金资助项目(NY2017030);南京邮电大学江苏省通信与网络技术工程研究中心开放课题资助项目(BK20181392)
作者简介:鲁 伟(1995-),男,硕士研究生,研究方向为边缘计算;宋荣方,教授,博导,研究方向为宽带无线通信。

服务器之间的距离,从而大大降低了用户设备的传输时延。在移动边缘计算系统中,任务卸载调度策略的好坏也会直接影响到系统时延和用户体验。终端将业务卸载至边缘计算服务器时,服务器通过优化业务调度顺序可以进一步降低时延和系统能耗。目前已有许多文献对 MEC 任务卸载调度进行了研究,MEC 卸载常见的衡量指标有时延、能耗以及时延和能耗综合权衡问题[6]。文献[7]研究了单用户单核服务器任务卸载情景,提出了一种基于李雅普诺夫优化的动态计算迁移算法,旨在优化应用的执行时延。文献[8]研究了单用户单核服务器任务卸载情景,提出了二进制粒子群优化算法,旨在优化系统的能耗。文献[9]利用流水车间调度模型对任务卸载调度进行建模,以交替最小化优化方法研究系统时延与能耗关系。文献[10]研究单用户多核任务卸载情景,利用遗传算法对单用户的能耗与时延关系进行了优化分析。受以上文献的启发,且目前多核多用户任务卸载研究较少,该文研究多核多用户任务卸载情景,采用混合流水车间模型进行建模,以最小化系统时延和能耗加权和为优化目标,采用模拟退火算法进行求解,通过仿真获得了多用户最优的任务卸载策略,最后对系统时延和能耗关系进行了分析。

1　系统模型与问题建模

该文研究多用户多核任务卸载情景。该边缘任务卸载系统包含了多个用户和一个多核的 MEC 服务器。每个用户之间卸载相互独立互不影响,每个用户的多个可卸载任务也相互独立互不影响。用户可以通过无线信道将任务上传至边缘服务器进行任务卸载。由于每个任务上传所需的时间和在核服务器上卸载的时间的不同,不合理的任务卸载顺序必将导致系统的总体时延较大,因此确定合理的任务卸载顺序至关重要。

1.1　任务调度与传输速率的定义

移动终端将各自的 N 项独立的计算任务卸载到 MEC[11]。记各自的任务集合为 $R=\{T_1,T_2,\cdots,T_N\}$,每个任务 T 用一对参数$\langle D_i,C_i\rangle$来表示,其中 D_i(bits)表示任务的数据量,C_i(cycles/bit)表示每比特的数据所需的计算资源。每个用户 N 个任务的卸载调度顺序定义为 $\sigma=\{\sigma_1,\sigma_2,\cdots,\sigma_N\}$,其中 T_{N_σ} 表示该任务 N 于第 i 次卸载到 MEC 服务器上。该文研究移动端配置单天线情景,一次只能发一个任务,任务 T_{N_σ} 的传输速率定义为:

$$R(p_i)=w\left(\log_2\left(1+\frac{g_0\ (L_0/L)^{\theta}P_i}{N_0 w}\right)\right) \tag{1}$$

其中,P_i 是任务 T_{N_σ} 的传输功率,g_0 是路径损耗常数,θ 是路径损耗指数,取值范围一般为 2 ~ 4,L_0 是参考距离,L 是终端与 MEC 服务器之间的距离,w 是系统带宽,N_0 是 MEC 服务器接收端的噪声功率谱密度。

1.2　系统时延和能耗模型

混合流水车间调度(hybrid flow-shop scheduling problem,HFSP)是一种车间作业排序问题[12]。如图 1 所示,设有 n 个独立的工件按照相同加工方向在 m 道工序上加工,m 道工序中至少有一道工序包含多台并行处理器[13]。

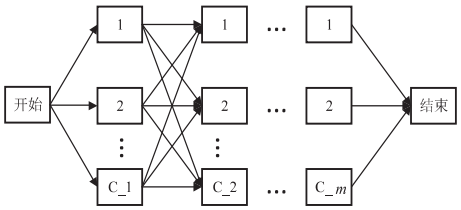

图 1　混合流水车间调度模型

模型一般满足以下条件:(1)同一阶段中所有机器都相同;(2)每个工件可以在某阶段的任意一台机器上进行加工;(3)任意时刻每个工件至多在一台机器上加工;(4)每台机器某时刻只能加工一个工件;(5)工件的加工过程不允许中断;(6)每台机器都有一个无限的存储空间。

在多用户多核服务器 MEC 系统中,可将卸载的任务看成是待加工的工件,每个计算任务都需要经过本地传输和服务器执行两道工序。在第一道工序中,移动设备负责任务的上传,在第二道工序中,MEC 服务器具有 M 个计算能力相同的处理器,因此可以利用混合流水车间模型对多用户多核服务器 MEC 系统的任务卸载调度进行建模。当任务 T_{N_σ} 卸载到 MEC 服务器上执行时,系统时延由三部分组成,即任务上传到服务器的时间 $t(1,\sigma_j)$、任务在服务器执行的时间 $t(2,\sigma_j)$ 和任务结果反馈到移动设备的时间,通常由于下行速率远远高于上行速率,因此可忽略结果的反馈时间。

$$t(1,\sigma_j)=\frac{d_{\sigma_j}}{R(p_{\sigma_j})} \tag{2}$$

$$t(2,\sigma_j)=\frac{d_{\sigma_j}\times c_{\sigma_j}}{f_{\text{ser}}} \tag{3}$$

$$\sigma_j\in(1,2,\cdots,N)$$

在多用户多核 MEC 系统中,每个用户完工时间定义为该用户最后一个任务在某个核上的完工时间[14],系统时延定义为每个用户最后一个任务在某个核上的完工时间的累加和,即:

$$\sum_i^k c_k(i_2,\sigma_N) \tag{4}$$

系统能耗定义为每个用户每个任务上传所消耗能耗的累加和,即:

$$E_{\text{up}} = \sum_{i=1}^{k}\sum_{j=1}^{N} t_k(1,\sigma_j) \times p_k(\sigma_j) \tag{5}$$

2 问题建模

基于以上分析,该文以最小化时延和能耗的加权和为目标,即:

$$p\ \min \sum_{i}^{k} c_k(i_2,\sigma_N) + \eta \times E_{\text{up}} \tag{6}$$

$$\text{s.t. } 0 \leqslant p_i \leqslant p_{\max}, i = 1,2,\cdots,N$$

$\sigma = \{\sigma_1,\sigma_2,\cdots,\sigma_N\}$,其中 $\sigma_i \in \{1,2,\cdots,N\}$, $i \neq j, \sigma_i \neq \sigma_j, \forall i,j,i,j = 1,2,\cdots,N$

其中,η 为权重因子,用于调节系统时延和能耗之间的数量级,当其较大时,表示对系统能耗的优化更加看重。该求解问题是一个优化问题,可以使用穷举算法遍历所有情况,但复杂度太高。考虑到模拟退火算法是一种借鉴于固体的退火原理的优化算法,计算过程简单,通用,鲁棒性强,适用于并行处理,可用于求解复杂的优化问题,所以用模拟退火算法对问题 p 进行求解。

3 模拟退火算法

模拟退火算法(simulated annealing algorithm, SAA)是一种基于蒙特卡洛迭代的随机寻优算法,其出发点是模仿物理中固定物质的退火过程与一般组合优化问题之间的相似性[15]。模拟退火算法在某一初温下,随着温度参数的不断下降,以一定的概率突跳,在解空间中随机寻找目标函数的全局最优解。为方便表示,将适应度函数 fitness 表示为目标函数值 E,目标函数值 E 越低,表示可行解越接近最优解。

$$E = \sum_{i}^{k} c_k(i_2,\sigma_N) + \eta \times E_{\text{up}} \tag{7}$$

算法流程:

(1)设定当前解:$T = T_0$,即开始退火的初始温度,随机生成一个初始解 $X_{\text{best}} = X_0$,并计算相应的目标函数值 $E(x_0)$,令 T 等于冷却进度表中的下一个温度值 T_i。

(2)产生新解与当前解的差值:对当前解 X_i 进行扰动,产生一个新解 X_{new},并计算相应的目标数值 $E(X_{\text{new}})$ 进而得到 $\Delta E = E(X_{\text{new}}) - E(X_i)$。

(3)判断新解能否被接受:若 $\Delta E < 0$, $X_{\text{best}} = X_{\text{new}}$,接受新解,否则新解 X_{new} 按照概率 $e^{-\frac{\Delta E}{T}} > \text{random}(0,1)$ 进行接受。

(4)更新温度,$T_{k+1} = \text{update}(T_k)$,在温度 T_{k+1} 下,再经过 k 次扰动和接受,即执行步骤 2 和步骤 3。

(5)找到可行解:判断 T 是否达到了终止温度,是,终止算法;否,则转到步骤 2 继续执行。

4 仿真结果与分析

下面对多用户多核服务器的 MEC 系统分别用基于混合流水车间模型的模拟退火算法(HFSP-SAA)与随机任务卸载(random task offload strategy, RTOS)的任务数与时延的关系任务卸载调度进行仿真并分析。仿真中计算任务的数据量 D_i 和所需的计算资源 C_i 都服从均匀分布,即 $D_i \sim \text{U}(2\ \text{davg}, 20\ \text{davg})$, $C_i \sim \text{U}(5\ \text{cavg}, 27.975\ \text{cavg})$,其中 davg = 1 kbit, cavg = 797.5 cycles/bit。表 1 列出了仿真所需要的参数及取值。

表 1 仿真参数与取值

参数	物理意义	取值
g_0	路径损耗常/dB	-40
L_0	参考距离/m	1
L	用户与 MEC 之间的距离/m	100
θ	路径损耗指数	4
w	传输带宽/MHz	3
N_0	噪声功率谱密度/(dBm/Hz)	-174
$p_{\max}$	最大传输功率/mW	100
p	传输功率(mW)	0.5、16、32
f_{ser}	MEC 服务器主频/GHZ	1
M	MEC 服务器的处理核数	1、2、3、4、5

图 2 展示了 $\eta = 0$ 时基于混合流水车间模型模拟退火算法在不同传输功率下 2 核 2 用户时延与卸载任务数的关系。

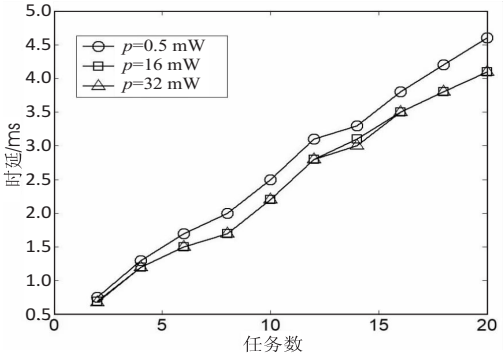

图 2 时延与卸载任务之间关系($M = 2$)

从图中可以看出,随着卸载任务数量的增大,时延呈现上升趋势。传输功率从 0.5 mW 增大至 16 mW,时延显著降低,但传输功率从 16 mW 增大至 32 mW,时延降低并不明显。

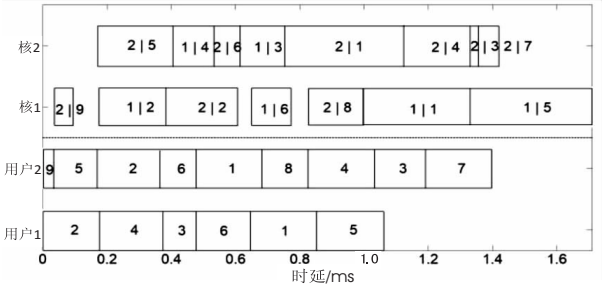

图 3 P = 0.5 mW 任务卸载甘特图

图 3 ~ 图 5 分别展示了在不同的传输功率下 2 核 2 用户的卸载任务调度的甘特图。用户的任务数字表示正在上传的任务序号,而核服务器上的数字表示该任务的归属,例如核 2 上数字的 2|5,表示核 2 正在处理用户 2 的第 5 个任务。

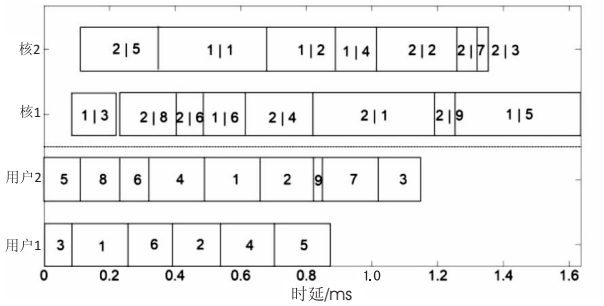

图 4　P = 16 mW 任务卸载甘特图

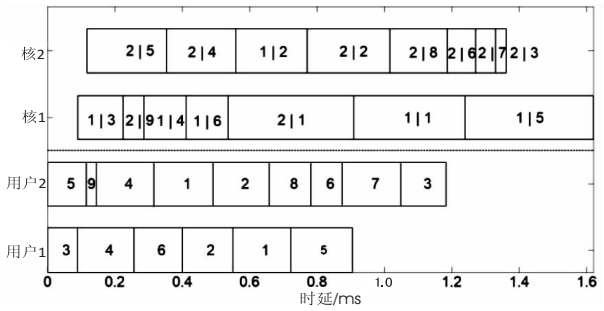

图 5　P = 32 mW 任务卸载甘特图

从甘特图中可以看出,当传输功率 P = 0.5 mW 时,核服务器一开始等待时间较长,任务上传时间过长,从而导致 MEC 服务器资源无法充分得到利用,且在处理任务过程中,因为传输功率低而导致核服务器有空闲等待的时刻,从而导致时延较高,且当卸载任务数量显著增大时,这种空闲等待情况更加明显,因此时延会显著增大。而当传输功率 P = 16 mW 时,任务上传时间减少,核服务器等待时间减少,且核服务器无空闲等待时刻,因此时延降低。当 P = 32 mW 时,尽管任务上传时间减短,但核服务器因为资源有限,上传的任务进入了缓存等待区域,因此传输功率的再次增大并没有换取时延的显著降低。

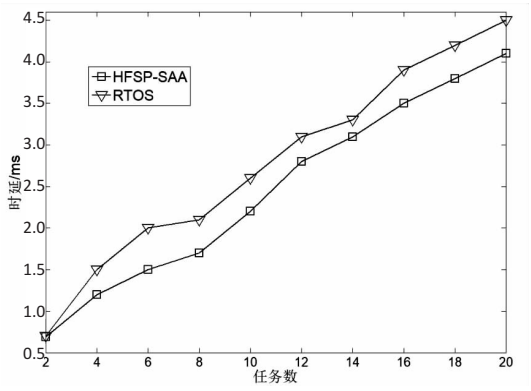

图 6　系统时延与卸载任务数量关系(P = 16 mW)

图 6 展现了基于混合流水车间模型的模拟退火算法(HFSP-SAA)与随机任务卸载(random task offload strategy,RTOS)的任务数与时延的关系。

从图中可以看出,随着任务数的增大,基于 HFSP-SAA 卸载策略比 RTOS 卸载策略的系统时延要少,这是因为 HFSP-SAA 卸载策略综合考虑了两道工序的加工时间,确定了合理的任务卸载顺序,从而使得系统时延得以减少,且随着任务数量的增大,HFSP-SAA 卸载策略的优势更加明显,因此提出的卸载策略可以有效降低时延,提高用户体验。

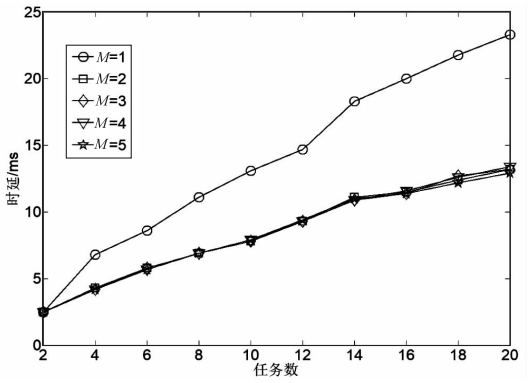

图 7　不同核数下时延与卸载任务数量关系

图 7 展示了在 2 用户不同核数情况下,系统时延与卸载任务数量的关系。从图中可以看出,当核数小于用户数时,系统时延优化瓶颈在核服务器等待和空闲时延上,此时核数的增加可以显著减少时延,而当核数大于或等于用户数时,核数的增加不会显著减少时延,系统时延优化瓶颈在第一道工序的任务上传上。由此可得出参与计算卸载最佳的核数应该等于或近似于参与任务卸载的用户数,由此可以实现服务器端能耗的节约,当参与调度的用户数改变时,动态调节核数,保证核数等于或近似于用户数时,从而可以有效降低用户任务卸载时延。

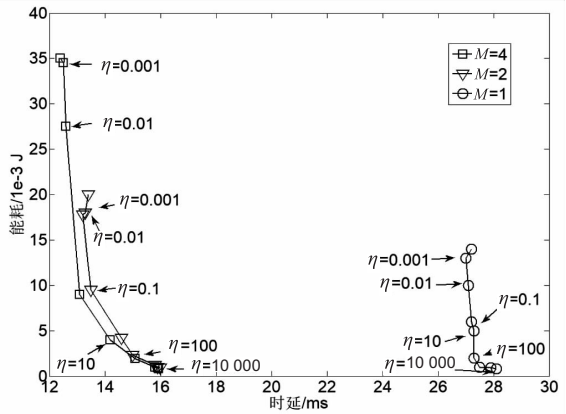

图 8　系统时延与能耗关系

图 8 展示了 2 用户情景下,不同核数不同的权重下,能耗与时延的优化关系。从图中可以看出,当用户数大于核数时,增加核数可以显著减少时延。系统时延随着 η 增大而增大,系统能耗随着 η 增大而减小,但

能耗呈现先陡峭后平缓减少的走势;陡峭部分的能耗说明当能耗增大到某一程度后,能耗的增加不会降低时延,当能耗小到某一程度,能耗与时延成反比关系,能耗降低会引起时延的增大。当 $M=1$ 时,用户数大于核服务器数量时,此时能耗的增加并不会引起时延降低,可取 $\eta=10\,000$,作为优化权重,从而实现节约能耗,对于用户数小于核数的 $M=4$ 和 $M=2$ 的情况,可取 $\eta=10$ 作为优化权重,此时能耗较低,时延较低,由此实现节约能耗的目的。

5 结束语

该文研究了多用户多核情景下多个独立任务调度和功率分配问题。基于混合流水车间调度模型和模拟退火算法,对系统时延和能耗进行加权和优化,获得了最佳的任务卸载调度甘特图。与随机任务卸载调度相比,提出的卸载调度策略时延较小。找到了一种基于混合车间模型的核服务器数量与参与调度的用户数的关系,从而确定最佳的核服务器数量,解决了当用户数大于核数时,系统时延的优化瓶颈。最后揭示时延与能耗之间的关系,根据核数与用户数关系,找到了不同情况下最佳的优化权重,从而达到了节约能耗的目的。

参考文献:

[1] GUBBI J, BUYYA R, MARUSIC S, et al. Internet of things (IoT): a vision, architectural elements, and future directions[J]. Future Generation Computer Systems, 2013, 29(7): 1645-1660.

[2] ARMBRUST M, FOX A, GRIFFITH R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.

[3] 王 妍,葛海波,冯安琪. 云辅助移动边缘计算中的计算卸载策略[J]. 计算机工程,2020,46(8):27-34.

[4] SATYANARAYANAN M. The emergence of edge computing[J]. Computer, 2017, 50(1): 30-39.

[5] ABBAS N, ZHANG Yan, TAHERKORDI A, et al. Mobile edge computing: a survey[J]. IEEE Internet of Things Journal, 2018, 5(1): 450-465.

[6] 张开元,桂小林,任德旺,等. 移动边缘网络中计算迁移与内容缓存研究综述[J]. 软件学报,2019,30(8):2491-2516.

[7] MAO Y, ZHANG J, LETAIEF K B. Dynamic computation offloading for mobile-edge computing with energy harvesting devices[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(12): 3590-3605.

[8] CAO S, TAO X, HOU Y, et al. An energy-optimal offloading algorithm of mobile computing based on HetNets[C]// International conference on connected vehicles & expo. Shenzhen, China: IEEE, 2016: 254-258.

[9] MAO Y, ZHANG J, LETAIEF K B. Joint task offloading scheduling and transmit power allocation for mobile-edge computing system[C]//IEEE wireless communications and networking conference. San Francisco, CA, USA: IEEE, 2017: 1-6.

[10] 凌雪延,王 鸿,宋荣方. 多核服务器边缘计算系统中任务卸载调度和功率分配的研究[J]. 南京邮电大学学报:自然科学版,2020,40(2):81-88.

[11] WANG K, YANG K, MAGURAWALAGE C S. Joint energy minimization and resource allocation in C-RAN with mobile cloud[J]. IEEE Transactions on Cloud Computing, 2018, 6(3): 760-770.

[12] 景会成,王 颖. 模拟退火算法优化 PSO-GA 算法解决柔性流水车间调度问题[J]. 小型微型计算机系统,2020,41(5):996-999.

[13] 杜利珍,王 震,柯善富,等. 混合流水车间调度问题的果蝇优化算法求解[J]. 中国机械工程,2019,30(12):1480-1485.

[14] 任彩乐,张超勇,孟磊磊,等. 基于改进候鸟优化算法的混合流水车间调度问题[J]. 计算机集成制造系统,2019,25(3):643-653.

[15] 韩晓辉,高 远,颜 丽,等. 基于模拟退火算法的电源规划[J]. 上海电力大学学报,2020,36(3):245-250.