



Cost-aware demand scheduling for delay tolerant applications



Xiumin Wang^{a,b,*}, Chau Yuen^c, Xiaoming Chen^d, Naveed Ul Hassan^e, Yiming Ouyang^a

^a School of Computer and Information, Hefei University of Technology, Hefei, China

^b The State Key Laboratory of Integrated Services Networks, Xidian University, China

^c Singapore University of Technology and Design, Singapore

^d Nanjing University of Aeronautics and Astronautics, Nanjing, China

^e Department of Electrical Engineering, LUMS School of Science and Engineering, Lahore, Pakistan

ARTICLE INFO

Article history:

Received 1 August 2014

Received in revised form

18 December 2014

Accepted 4 April 2015

Available online 23 April 2015

Keywords:

Demand scheduling

Peak resource reduction

Completion time minimization

ABSTRACT

In this paper, we study the problem of demand scheduling for delay tolerant applications. With regard to the time-varying resource cost per unit size of the demand, we study two optimization problems: (1) how to minimize the peak resource usage, while making sure that each demand is served before the deadline; (2) how to minimize the longest completion time of all the demands under a given maximum allowable resource constraint. For the first problem, we prove that it is NP-hard, under the general setting that the demands are of different sizes and require several continuous time slots to complete. We then provide an integer linear programming solution, and propose an efficient heuristic algorithm. For a special case of the same size demand and single serving time slot, the proposed algorithm is proved to be optimal. We further study a special case that all the demands have the same deadline, and prove that the proposed algorithm can achieve three times the optimal solution if the number of serving time slots required for each demand is at most two. For the second problem, we also prove it to be NP-hard and formulate it into an integer linear programming. An efficient polynomial-time algorithm is then proposed, whose completion time is proved to be at most two times the optimal minimum completion time under a specific setting. Finally, simulation results demonstrate the superiorities of the proposed schemes.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In this paper, we study two scheduling problems for delay tolerant applications (Laoutaris et al., 2013; Krithivasan and Iyer, 2005; Jiang et al., 2014; Shin et al., 2012; Yoon et al., 2014): (1) how to minimize the peak resource usage for a given set of demands and time-varying resource cost while at the same time making sure that each demand can be served without missing their deadlines; (2) how to complete the demands in the earliest time slot for a given maximum allowable resource at each time slot.

The above problems occur in many scenarios such as in communication network, heat management, smart grid systems as shown in Table 1 (Marcon et al., 2010; Malandrino et al., 2012; Stanojevic et al., 2010; Ghanem et al., 2007; Chhabra et al., 2010; Luo et al., 2013; Subramanian et al., 2013; Hassan et al., 2013a, 2013b; Seiden et al., 2000; Phillips et al., 1998; Lua et al., 2003;

Hsu et al., 2011). For example, in communication network, there are multiple demands with different throughput requirements and the throughput per channel at different time slots may be different due to network congestion, etc. Under such scenario, there are two problems that should be considered. Firstly, if each demand has to be served by a deadline, to reduce the resource usage (i.e., the number of channels), we need to decide when to schedule each demand such that the peak number of channels used at each time slot is minimum while all the demands are served without missing their deadlines. Secondly, if the maximum number of channels available at each time slot is constrained, shortening the completion time of all the demands is an attractive feature to improve the QoS performance.

The second scenario is the heat management for computing jobs or data center (Yang et al., 2008; Wang et al., 2009; Zhou et al., 2010). It is noted that the effectiveness and lifetime of a computer system is directly related to its operating temperature. In general, the cooling efficiency is time-varying, e.g. it may depend on outdoor temperature, electricity cost, etc. To maintain a steady temperature for the server room, we may want to reschedule the job accordingly. However, it might miss the deadlines of the jobs. Thus, one problem is how to minimize the

* Corresponding author.

E-mail addresses: wxiumin@hfut.edu.cn (X. Wang), yuenchau@sutd.edu.sg (C. Yuen), chenxiaoming@nuaa.edu.cn (X. Chen), naveed.hassan@lums.edu.pk (N.U. Hassan), oyymbox@163.com (Y. Ouyang).

Table 1
Application scenarios.

Parameters	Communication network	Heat management	Smart grid
Demand	Throughput	Computing jobs	Energy demand
Time-varying parameters	Channel bandwidth	Cooling efficiency	Energy cost
Resource	Number of channels	Temperature	Budget

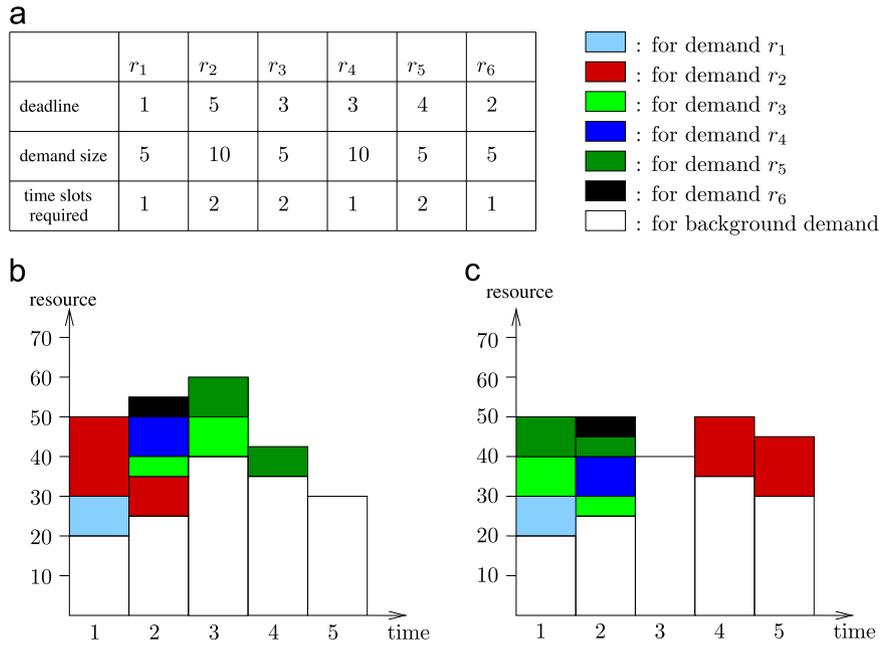


Fig. 1. An illustration of different scheduling schemes with unit resource cost 2, 1, 2, 1.5, 1.5 at time slots 1, 2, 3, 4, 5 respectively.

highest temperature and at the same time make sure that each job is scheduled without missing the deadline. In addition, the maximum allowable temperature may be predefined in the system and cannot be violated during job processing. Under such setting, an important work is to complete all the jobs as early as possible without violating the predefined allowable temperature.

The last example we want to show is the energy management in smart grid, where the energy cost is time-varying because of the renewable energy sources or the dynamic energy market (Hassan et al., 2013a,b; Tsitsiklis and Xu, 2012; Liu et al., 2014). Due to the constraint of budget available, we need to determine when to serve the demands such that the completion time of the energy demands is minimum while the maximum budget allocated for the energy cost is not violated.

In the literature, to minimize the peak resource, one approach is to “shift” the delay-tolerant demands from the busiest hours to the idle hours (Marcon et al., 2010; Malandrino et al., 2012; Hassan et al., 2013b). Specifically, the work in Malandrino et al. (2012) proactively “pushes” the content to the users before or no later than they request it, by exploiting the predictability of the future demands. As such, the peak traffic load at each time slot can be minimized. Completion time minimization problem also has been studied in the literatures (Phillips et al., 1998; Lua et al., 2003; Hsu et al., 2011). For example, the work in Lua et al. (2003) schedules jobs online and proposes a 2-competitive algorithm. However, all previous works either assume that the demands to be considered have the same size or the demands can be completed within single time slot. Different from previous works, in this paper, we consider a more general model, where the demands are of different sizes and may need several continuous time slots to complete. In

Table 2
Main notations and their descriptions.

B_t	The resource required to serve the unit size of the demand at t
c_i	The size of demand r_i that should be served at each serving time slot
$\hat{c}_{i,t}$	the variant to denote the size of demand r_i served at time t
d_t	The variant to denote the total resource used at time t
R	The set of all the demands in the system
r_i	The i -th demand in R
Δt_i	The number of continuous time slots required to complete demand r_i
T_i	The deadline of demand r_i
\hat{t}_i	The variant to denote the starting time slot of demand r_i
γ_t	The resource used for the background demands at time t

addition, the resource required to serve the unit size of the demand (we call it *unit resource cost*) is time-varying. For example, in data centers, unit energy price/cooling cost may fluctuates over times (Tsitsiklis and Xu, 2012), depending on the environment condition and electricity market.

Take Fig. 1(a) as an example, where we consider the first scheduling problem, i.e., minimizing the peak resource usage. Here, the resource can represent all the parameters (e.g., channels, temperature) listed in Table 1. Assume that there are six demands $\{r_1, r_2, r_3, r_4, r_5, r_6\}$ to be scheduled and some background demands that must be served in some fixed times. The numbers of continuous time slots required to serve these six demands are 1, 2, 2, 1, 2, 1 respectively, and at each serving time slot, the sizes of demands $r_1, r_2, r_3, r_4, r_5, r_6$ are 5, 10, 5, 10, 5, 5 respectively. Suppose that the demands $r_1, r_2, r_3, r_4, r_5, r_6$ are required to be completed by time slot 1, 5, 3, 3, 4, 2 and the unit resource cost at time slots 1–5 are

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات