



Bottleneck-Aware Resource Allocation for Service Processes: A New Max-Min Approach

Shoulu Hou, Beijing Information Science and Technology University, China

Wei Ni, CSIRO, Australia

 <https://orcid.org/0000-0003-0780-4637>


Ming Wang, University of International Business and Economics, China

 <https://orcid.org/0000-0002-6641-3700>

Xiulei Liu, Beijing Information Science and Technology University, China

Qiang Tong, Beijing Information Science and Technology University, China

Shiping Chen, CSIRO, Australia

 <https://orcid.org/0000-0002-4603-0024>

ABSTRACT

In 5G systems and beyond, traditional generic service models are no longer appropriate for highly customized and intelligent services. The process of reinventing service models involves allocating available resources, where the performance of service processes is determined by the activity node with the lowest service rate. This paper proposes a new bottleneck-aware resource allocation approach by formulating the resource allocation as a max-min problem. The approach can allocate resources proportional to the workload of each activity, which can guarantee that the service rates of activities within a process are equal or close-to-equal. Based on the business process simulator (i.e., BIMP) simulation results show that the approach is able to reduce the average cycle time and improve resource utilization, as compared to existing alternatives. The results also show that the approach can effectively mitigate the impact of bottleneck activity on the performance of service processes.

KEYWORDS

Bottleneck-Aware, Equal or Close-to-Equal Service Rate, Max-Min Problem, Resource Allocation, Service Process

INTRODUCTION

In 5G and beyond (B5G) systems, the emerging new service scenarios and business models driven by user demands will differ from today's in terms of deepened connections and extended service chains (Yu et al., 2020; Chien et al., 2019; Hou et al., 2017). Traditional generic service models are no longer appropriate for highly customized and intelligent services (Aazam et al., 2019; Van Hee et al., 2001). Diversity and dynamics are the new features of user demands, which require the service providers to reinvent their service models in accordance with fast-changing demands. For example, Figure 1 shows a change occurring in the service process in the tourism industries during and after

DOI: 10.4018/IJWSR.2021070101

the coronavirus disease 2019 (COVID-19) era. To protect public health, a new regulation requests that tourists send both health code and journey data within 14 days to the travel agencies (Ienca et al., 2020). To obtain the health information of each applicant, a service designer needs to add two new operations “Check Health Code” and “Check Journey Data” after the operation “Receive Inquiry”; see the dotted box in Figure 1. Reinventing service models involves allocating resources to the newly added tasks to rapidly adapt to market changes (Van Hee et al., 2001; Mendling et al., 2018). Additionally, a sudden burst of user requests requires the service providers to add resources to accommodate the increased load conditions, e.g., the newly initiated service requests for insurance claim during or after the occurrence of disasters (Doan et al., 2019; van der Aalst et al., 2007). This also involves the problem of allocating newly added resources to pending tasks to respond to user service requests.

It is not easy to develop a simple and effective approach to allocating resources globally optimally to quickly satisfy new business service requirements in future B5G era. The reason is that adding resources to task nodes with no growth in others does not improve performance, e.g., the number of users that can be served per unit time and the average response time per request. Generally, a service consists of one or more business processes, whose activities work together to deliver the specific service according to internal business rules. A business process model defines the activities to be executed, their data objects and resources, and the execution order of activities (Combi et al., 2009; Natschläger et al., 2015). Each activity has the important attribute, i.e., duration, which specifies the allowed temporal spans of the activity (Combi et al., 2009), and requires a certain number of resources to complete. The appropriate allocation of resources to an activity has a direct impact on the performance of entire service processes, e.g., the cycle time of process instances (Sheng et al., 2009) and the data quality of information systems (Liu et al., 2020). However, the optimal resource allocation under constraints, as a typical problem of operations research (OR), is NP-hard and difficult to solve (Doan et al., 2019; Xu et al., 2019; Ma et al., 2020).

To meet the required quality of service (QoS), some existing studies have attempted to address the resource allocation under unlimited resources by resorting to cloud or crowdsourcing platforms. These platforms serve as resource providers (Bessai et al., 2016; Halima et al., 2017; Rosinosky et al., 2016). The rich resources of the Internet bring service users powerful computing power, as well as problems of sensitive information disclosure, especially for tasks with sensitive, confidential, or classified data in government agencies, banking, and healthcare systems. Therefore, these studies are only applicable to tasks with non-sensitive data, e.g., the service supervision process of French telecoms operator (Halima et al., 2017). Other existing studies tried to address the resource allocation under limited resources (He et al., 2016; Van Hee et al., 2001; Boxma et al., 1990; Sheng et al., 2009; Liu et al., 2020). These studies are applicable to service processes involving the generation of sensitive data, in which tasks can only be served by specified government officials and local mechanisms, e.g., the approval process of urban-poverty relief (Hu et al., 2016). However, the studies adopted different criteria to guide allocation actions, e.g., data quality and resource priority, which are difficult to reach the optimal throughout performance due to a shortage of available resources and unbalanced allocation of limited resources.

It is nontrivial to find an optimization criterion to detect the bottleneck that affects service performance, improve allocation results, and promote service ability. This is because an unbalanced allocation leads to the occurrence of bottleneck activity and waste of resources (Huang et al., 2011; Wang et al., 2019). Specifically, the activity with the lowest service rate of all activities is the bottleneck which affects the overall performance of service processes, especially under a heavy load condition. A resource-activity allocation criterion is imperative for a process model that indicates what the best resource allocation decision is.

This paper presents a new bottleneck-aware resource allocation approach for service processes accompanied by generations of sensitive data. The approach can balance the service rates of different activities from a global view. This is achieved by finding the optimal allocation that maximizes the minimum service rate of activities within a service process. The key contributions of the paper can be summarized as follows.

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/article/bottleneck-aware-resource-allocation-for-service-processes/285925?camid=4v1

This title is available in e-Journal Collection, Computer Science and IT Knowledge Solutions e-Journal Collection, Business, Administration, and Management e-Journal Collection, Computer Science, Security, and Information Technology e-Journal Collection, Digital Marketing, E-Business, and E-Services Collection - e-Journals, Networking, Mobile Applications, and Web Technologies Collection - e-Journals. Recommend this product to your librarian:

www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Ontology Driven Data Mediation in Web Services

Meenakshi Nagarajan, Kunal Verma, Amit P. Sheth and John A. Miller (2007). *International Journal of Web Services Research* (pp. 104-126).

www.igi-global.com/article/ontology-driven-data-mediation-web/3111?camid=4v1a

A Metamorphic Relation-Based Approach to Testing Web Services Without Oracles

Chang-ai Sun, Guan Wang, Baohong Mu, Huai Liu, ZhaoShun Wang and T. Y. Chen (2012). *International Journal of Web Services Research* (pp. 51-73).

www.igi-global.com/article/metamorphic-relation-based-approach-testing/64223?camid=4v1a

Big Data Mining Using Collaborative Filtering

Anu Saini (2019). *Web Services: Concepts, Methodologies, Tools, and Applications* (pp. 702-711).

www.igi-global.com/chapter/big-data-mining-using-collaborative-filtering/217858?camid=4v1a

Runtime Reusable Weaving Model for Cloud Services Using Aspect-Oriented Programming: The Security-Related Aspect

Anas M.R. Alsobeh, Aws Abed Al Raheem Magableh and Emad M. AlSukhni (2018). *International Journal of Web Services Research* (pp. 71-88).

www.igi-global.com/article/runtime-reusable-weaving-model-for-cloud-services-using-aspect-oriented-programming/193862?camid=4v1a