Accepted Manuscript

Title: Fast Multi-resource Allocation with Patterns in Large Scale Cloud Data Center

Author: Jiyuan Shi Junzhou Luo Fang Dong Jiahui Jin Jun Shen

PII: S1877-7503(17)30518-5
DOI: http://dx.doi.org/doi:10.1016/j.jocs.2017.05.005
Reference: JOCS 674

To appear in:

Received date: 17-1-2017
Revised date: 5-4-2017
Accepted date: 5-5-2017

Please cite this article as: Jiyuan Shi, Junzhou Luo, Fang Dong, Jiahui Jin, Jun Shen, Fast Multi-resource Allocation with Patterns in Large Scale Cloud Data Center, <![CDATA[*Journal of Computational Science*]]> (2017), http://dx.doi.org/10.1016/j.jocs.2017.05.005

Fast Multi-resource Allocation with Patterns in Large Scale Cloud Data Center

Jiyuan Shi[a], Junzhou Luo[a,*], Fang Dong[a], Jiahui Jin[a], Jun Shen[b]

[a] School of Computer Science and Engineering, Southeast University, Nanjing, P.R. China
[b] School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia

Abstract

How to achieve fast and efficient resource allocation is an important optimization problem of resource management in cloud data center. On one hand, in order to ensure the user experience of resource requesting, the system has to achieve fast resource allocation to timely process resource requests; on the other hand, in order to ensure the efficiency of resource allocation, how to allocate multi-dimensional resource requests to servers needs to be optimized, such that server's resource utilization can be improved. However, most of existing approaches focus on finding out the mapping of each specific resource request to each specific server. This makes the complexity of resource allocation problem increases with the size of data center. Thus, these approaches cannot achieve fast and efficient resource allocation for large-scale data center. To address this problem, we propose a pattern based resource allocation mechanism based on the following findings. In a real-world cloud environment, the resource requests are usually classified into limited types. Thus, the mechanism first utilizes this feature to generate pattern information, which indicates which types of resource requests are suitable to be allocated together to a server. Then, the mechanism uses the pattern information as guidelines to make fast resource allocation decision and fully utilize server's multidimensional resources. Simulation experiments based on real and synthetic traces have shown that our mechanism significantly improves system's resource utilization

[☆] Fully documented templates are available in the elsarticle package on CTAN.

[*] Corresponding author

Email addresses: `jiyuanshi@seu.edu.cn` (Jiyuan Shi), `jluo@seu.edu.cn` (Junzhou Luo), `fdong@seu.edu.cn` (Fang Dong), `jjin@seu.edu.cn` (Jiahui Jin), `jshen@uow.edu.au` (Jun Shen)