Accepted Manuscript

Performance measurement of data flow processing employing software defined architecture

Lei Liu, Yongjian Ren, Lizhen Cui, Yuliang Shi, Qingzhong Li

PII: S0167-739X(17)31718-1
DOI: https://doi.org/10.1016/j.future.2017.12.050
Reference: FUTURE 3888

To appear in: *Future Generation Computer Systems*

Received date : 31 July 2017
Revised date : 25 December 2017
Accepted date : 26 December 2017

Performance Measurement of Data Flow Processing Employing Software Defined Architecture

Lei Liu[a], Yongjian Ren[a], Lizhen Cui[a], Yuliang Shi[a], Qingzhong Li[a]

[a]*No.1500 Shunhua Road, Jinan, China*

Abstract

With the development of information technology, the importance of big data is quickly highlighted. Big data applications show great value to individuals, companies and governments. Recently, researches on the storage and utilization of big data have achieved considerable results. The prosperity of big data applications is a thrust of drawing attention to the system performance such as timeliness, computational and communication resources. Data retransmission caused by the violation of the stringent delay bound may result in the reprocessing of these data, which would have a negative effect on user experience. To fill this gap, a software defined architecture is developed in this work so that the appropriate start point of processing can be found for the data need to be reprocessed. For further improvement of the processing performance, two models are presented to this software defined architecture. In the optimized model, a priority queue is employed to facilitate the processing efficiency. In addition, data flows transmitting through networks exhibit obvious self-similar characteristics. Performance analysis without taking traffic self-similarity into account may lead to unexpected results. In the optimized model, the tightly coupled system makes performance analysis difficult. Therefore, a decomposition approach is employed to divide the coupled system into a group of single server single queue systems. Finally, the developed model is validated through extensive experimental results.

Keywords: software defined architecture; performance measurement; data flow processing; queueing system; self-similar.

1. Introduction

Big data [1, 2, 3, 4] creates values for business and society, but poses challenges in terms of networking management and analytics [5]. Big data is not just characterized by Volume. It is also depicted by Variety, Velocity, Value, and Veracity. Big data is not just data. It also means new ideas and methods. The value of big data can only be revealed after effective mining and analysis. Big data has high predictive power. In traffic management, business transactions, weather forecasting, national security, mobile services, manufacturing and other areas, big data is playing a significant role on improving efficiency, saving resources and optimizing decision-making [6, 7, 8, 9, 10, 11]. Data flow processing of big data applications usually includes acquisition, preprocessing, storage, analysis and mining. The effective utilization of big data may reduce the operating cost of enterprises and society and improve efficiency. In big data era, data flow processing is a challenging task because of the variety and velocity characteristics of big data [12, 13, 14, 15]. It should be well fault-tolerant, flexible, and extensible [16, 17]. A large number of practical applications put forward higher requirements on big data analysis, such as the timeliness, computational and communication resources [18, 19, 20, 21, 22]. For example, according to an estimate, in 2014 there were 9 billion interconnected things in China and it is expected to reach 24 billion by 2020 [23]. The rapid increase of devices could exhaust current communication resources while the data created by these devices could bring great pressure on computational resources. In the fields of weather forecasting and traffic management, people not only want to make better decisions by taking advantage of big data, but also have strict demands on response time. We note that analysis of big data tends to be divided into phases. When the violation of the stringent delay bound occurs, data may be retransmitted and reprocessed [24]. If fault-tolerant mechanisms are not optimized, there could be some unnecessary duplicates of processing stages [25]. Therefore, in this paper, the software defined architecture in which a controller is employed is developed to optimize the performance of data reprocessing. When data retransmission occurs, the controller should give a proper starting point to avoid redundant processing stages. For example, when handling a business at a bank