Going Fast and Fair: Latency Optimization for Cloud-Based Service Chains

Yuchao Zhang, Ke Xu, Haiyang Wang, Qi Li, Tong Li, and Xuan Cao

Abstract

State-of-the-art microservices have been attracting more attention in recent years. A broad spectrum of online interactive applications are now programmed to service chains on the cloud, seeking better system scalability and lower operating costs. Different from the conventional batch jobs, most of these applications consist of multiple stand-alone services that communicate with each other. These step-by-step operations unavoidably introduce higher latency to the delay-sensitive chained services.

In this article, we aim at designing an optimization approach for reducing the latency of chained services. Specifically, presenting the measurement and analysis of chained services on Baidu's cloud platform, our real-world trace indicates that these chained services are suffering from significantly high latency because they are mostly handled by different queues on cloud servers for multiple times. However, such a unique feature introduces significant challenges to optimize a microservice's overall queueing delay. To address this problem, we propose a delay-guaranteed approach to accelerate the overall queueing of chained services while obtaining fairness across all the workloads. Our evaluations on Baidu servers shows that the proposed design can successfully reduce the latency of chained services by 35 percent with minimal impact on other workloads.

Introduction

The rapid growth of service chains is changing the landscape of cloud-based applications. Different stand-alone components are now handled by cloud servers, providing cost efficient and reliable services to Internet users. It is known that the workloads from service chains are more complex than the traditional non-interactive (or batch) workloads [1]. Non-interactive workloads are the workloads that can be processed on only one specific server and do not need interactions with other servers (such as scientific computing and image processing). Being not strictly time-sensitive, these workloads can be scheduled to run anytime as long as they can be finished before a soft deadline, while interactive workloads from service chains are the workloads that have to go through multiple servers to apply different functions (such as business transactional and complex gaming control), and these chained services typically process real-time user requests. However, the interactions unavoidably introduce additional latency, making the performance for service chains in urgent need to be ensured.

We then measure the interactive workloads performance on Baidu's cloud platform. The real-world traces indicate that interactive workloads are really suffering from significantly longer latency than non-interactive workloads. The measured case is shown in Fig. 1a, where *Nuomi* is a group buying application, *Waimai* is a take-out service, and *Alipay* is an online payment platform. When a user clicks an item on Nuomi, the latency is quite short because this query does not require many interactions among services. However, the story will be different when this user orders a take-out and purchases the item. In this case, the request goes through Nuomi, Waimai, and then Alipay. In other words, this interactive workload consists of several highly-dependent operations that have to be processed on different servers separately. As shown in Fig. 1b, there are six procedures for interactive workloads and only two procedures for non-interactive workloads. It is easy to see that such interactive workloads in chained applications will introduce extra latency to users because these requests will be handled by different services for multiple times.

Unfortunately, we find that most existing workload scheduling approaches are designed to re-schedule [2] and leverage different priorities [3, 4] on individual queues. In other words, these optimizations are made on intermediate servers separately, so the overall latency of interactive workloads is still unpredictable. To better optimize the overall latency of chained services, we apply a latency estimation approach to predict overall latency and try to accelerate the interactive workloads. Furthermore, we design a feedback scheme to ensure workload fairness and avoid remarkable degradation of non-interactive workloads. Our real-world deployments on Baidu indicate that the proposed Delay-Guarantee (D^3G) framework can successfully reduce the latency of interactive applications with minimal impact on other workloads.

The main contributions of this article are summarized as follows:

- We present a measurement and latency analysis of service chains in Baidu networks and disclose the long latency of interactive workloads.
- We design the D^3G algorithm to accelerate interactive workloads in a global manner other than in each independent server, and leverage a latency estimation algorithm and a feedback scheme to ensure fairness.
- We evaluate our methods on servers in Baidu networks, and the extensive experiment results show that D^3G succeeds in accelerating interactive chained applications while ensuring workload fairness.

Yuchao Zhang is with Beijing University of Posts and Telecommunications.

Ke Xu is with Tsinghua University.

Haiyang Wang is with University of Minnesota at Duluth.

Qi Li is with Graduate School at Shenzhen, Tsinghua University.

Tong Li is with Huawei.

Xuan Cao is with Baidu.

Digital Object Identifier: 10.1109/MNET.2017.1700275

0890-8044/17/$25.00 © 2017 IEEE

Background

As more applications are deployed on clouds for better system scalability and lower operating costs, service chains are developing quickly. Many studies have shown that latency is particularly problematic when interaction latency occurs together with network delays [5]. In this section, we will first introduce the related work about service chains and interactive applications, and then present the measurements from a real-world trace, which motivates this article.

Related Work

To optimize cloud-based applications, many researchers focused on minimizing latency, which do advance the state-of-art. We classify this literatures into two categories. First, some literature focused on network processing latency. For example, Webb *et al*. [6] proposed a nearest server assignment to reduce client-server latency. Vik [7] explored the spanning tree problems in a distributed interactive application system for latency reduction. The authors in [8] and [9] introduced game theory into this topic and modeled the latency problem in datacenters as a bargaining game. Second, some research aims at reducing service latency. Web service is the most related, which is an effective mechanism for data and service integration on the Web. Some studies succeeded in dissecting latency contributors [10], showing that back-office traffic accounts for a significant fraction of Web transaction latency in terms of requests and responses [11].

The above studies handle different components of overall latency, but they ignore the effects brought by service chains, e.g., the case in Fig. 1. When service 1 receives an interactive request (TCP:0) that cannot be served in service 1 (i.e., this request requires data from services 2 and 3), then service 1 will make a new TCP connection (TCP:1) to service 2, and the original connection (TCP:0) would be hung up. In the same way, service 2 will make a new TCP connection (TCP:2) to service 3. When service 2 receives the corresponding results from service 3, the TCP:2 connection would be released and the TCP:1 connection recovered. After service 2 sends the results back to service 1, TCP:1 connection would be released and the original TCP:0 connection recovered. Thus, the interactive workloads are suffering from longer latency due to the interactions among multiple intermediate servers. Many researchers investigated the latency for these applications, and their research showed that although these applications are quite delay-sensitive, service performance is greatly affected by interactions. To address this problem, some studies suggested that the interactions of different services should be further dissected to better understand the performance implications [11], and some researchers have already begun to pay attention to interaction latency [12].

Our study explores the potential to reduce the response time for service chains and guarantee the non-interactive workloads simultaneously. In particular, we accelerate the interactive workloads by building a new dedicated queue and trying to adjust resource allocation among different queues. By leveraging a feedback scheme, we can bound the influence on non-interactive workloads. We'll describe the algorithm in detail after introducing our motivation below.

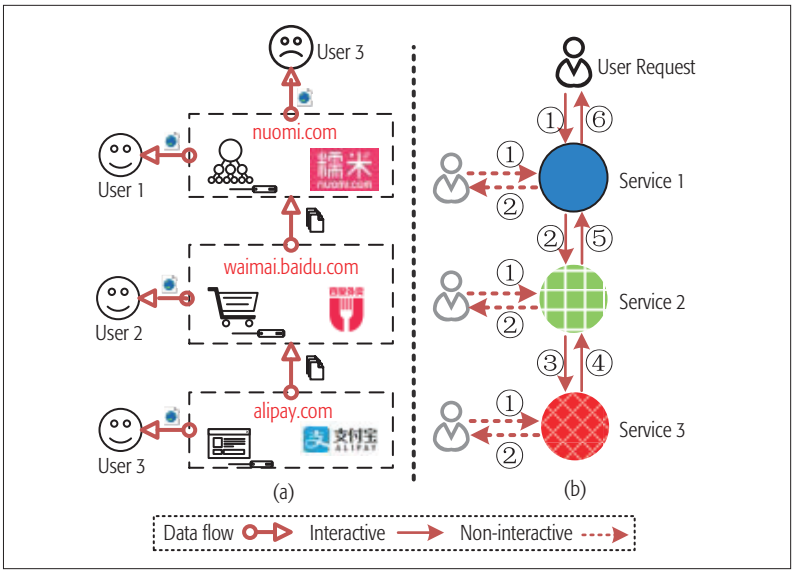

FIGURE 1. The processes of interactive and non-interactive workloads.

Measurement and Motivation

In this section, we conduct measurements in Baidu networks and disclose the long latency of chained services. This article was motivated by the goal to accelerat interactive workloads while not affecting non-interactive workloads (to ensure fairness).

As the largest Chinese search engine, Baidu has dozens of applications deployed in its networks. These applications cover every corner of people's lives, and they can further cooperate with each other to provide more comprehensive functions (as shown in the introduction section).

To evaluate the performance of these services, we measured the workload latency from one server cluster at Baidu. In particular, we monitor all the workloads, record the response time of service calls, and then calculate the average latency per minute of both interactive and non-interactive workloads by analyzing the trace log. We grab the log of these two kinds of workloads from 0:00 to 24:00 on April 7, 2016, and draw the statistical figures in Fig. 2. The x-coordinate denotes the time in one day while the y-coordinate denotes the service latency in ms. From these results, we come to the following conclusions:

- The average latency of non-interactive workloads is about 60 ms to 70 ms, while that of the interactive workloads is nearly 500 ms, that is, the interactive workloads are suffering from seven times longer latency compared to the non-interactive workloads.
- Even when the network is not congested (e.g., during the night), the interactive workload latency is still much longer than the non-interactive workload latency.
- When there is a slight burst, for example, at 11:00 am or 16:00 pm, the performance for interactive workloads is obviously influenced, making latency even higher.

As we analyzed before, non-interactive workloads can be completed in just one instance while the interactive workloads have to go through differ-

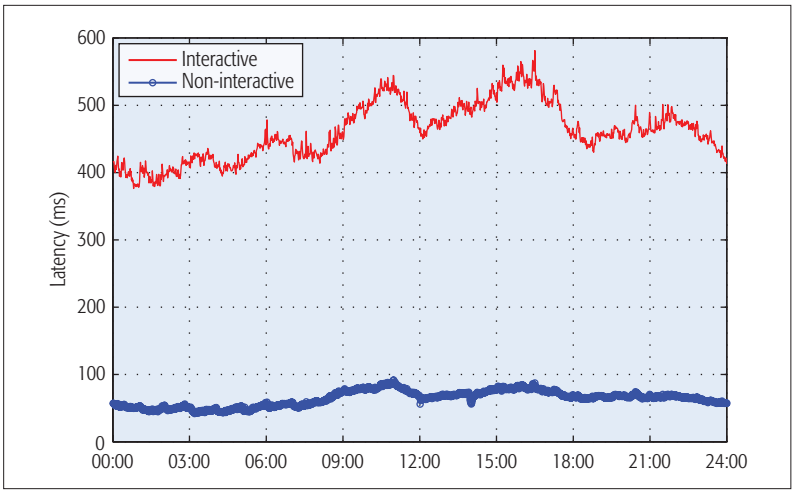

FIGURE 2. The workload latency in Baidu networks.

ent servers one after another. To optimize the performance of delay-sensitive interactive workloads, we should accelerate the processing of these workloads, such as assigning higher priorities or allocating more resources. However, improving interactive workloads will unavoidably affect non-interactive workloads because they are sharing the same infrastructures. Hence a fair optimization scheme should have the following two characteristics: reduce latency for delay-sensitive interactive workloads, and ensure fairness across all workloads (not to degrade non-interactive workloads severely).

Dynamic Differentiated Service with Delay-Guarantee

In this section, we study the essence of the latency gap before introducing the design philosophy of our approach. According to the philosophy, we design an algorithm called the Dynamic Differentiated service with Delay-Guarantee (D^3G), which reduces the latency of service chains while ensuring workload fairness.

The Components of Latency

As interactive applications consist of basic functions applied on different servers, those workloads should go through multiple servers in a specific order so that the required functions are applied step by step. Therefore, the interactive workloads will be queued in servers several times while the non-interactive workloads will be queued just once.

To be specific, we analyze the latency of interactive $R_{i,j}$ and non-interactive $R_{i,i}$ workloads. As interactive workloads travel across multiple servers and are queued in each one, the final latency is the sum of the queuing and serving time on each server. The transfer time among different servers also contributes to latency. Non-interactive workloads only go through one particular server with only one queuing and serving time. Thus, the overall latency of interactive workloads is much longer than that of non-interactive workloads (Fig. 2).

Design Philosophy

As the patience of users is limited and they would abandon the system once latency exceeds their patience, interactive workload latency is essential for these delay-sensitive applications. This limited patience can be formulated as an exponential function [13]. Considering that one exponential function is not precise enough to model user patience, we use a weighted sum of exponential functions to calculate user patience in the D^3G algorithm. With this expected patience, the system would have a leaving rate, rephrased as follows: when overall latency exceeds user patience, users will abandon the system, and this leads to an abandoning rate of waiting queues.

To ensure constant service and prevent users from abandoning the system, interactive workloads should be scheduled within user tolerance. To do so, we do re-scheduling and resource adjustment in this work. Specifically, we separate interactive workloads from non-interactive workloads and make them pending in different queues. We leverage two queues in each server. Q_l represents the queue for non-interactive requests, and Q_r represents the queue for interactive requests. The two queues share the infrastructure and resources in the same server. The difficulty in accelerating interactive workloads is that allocating more resources to Q_r will unavoidably affect the process of Q_l. Thus, how to share the resources on one server among different workloads becomes a key concern.

To address this issue, we design D^3G, which can adjust resource allocation among different kinds of workloads automatically and in real time. To make D^3G more intelligent, we design an estimation algorithm to pre-calculate the processing time on other servers. Furthermore, we also introduce a feedback scheme to reduce the negative impact on non-interactive workloads.

D^3G Framework

As we described in the previous subsection, we separate interactive workloads from non-interactive workloads and make them queued independently. We design a latency estimation algorithm, and once the estimated latency exceeds user patience, we dynamically adjust the resource allocation among queues according to a feedback scheme. Thus, the interactive workloads will be accelerated in all intermediate servers and finally enjoy a latency level that is comparable to the non-interactive workloads.

The framework of D^3G is shown in Fig. 3. For a specific server, it receives interactive workloads from other servers, and at the same time receives both interactive and non-interactive requests from users. A request type matching scheme will check whether this request should enqueue in Q_r (for interactive workloads) or in Q_l (for non-interactive workloads), with source (s), destination (d), and function (f). For each queue, the latency estimating algorithm pre-calculates the overall latency of these requests. If the estimated latency for interactive workloads exceeds user patience, the resource adjustment module would allocate the more resources to the interactive queue Q_r and update the request processing module. Then the processing speed of interactive workloads is thus promoted. This process executes in real-time automatically.

In the latency estimation algorithm, when a request with $\langle s, d, f \rangle$ enters a queue, we update the queue information and record the arriving

time. Once a request begins to be served, we record the beginning time and the queueing time. If this request is an interactive one, it will be transmitted to the next service. If this request is a non-interactive one, we can get the finish time directly, which is calculated by summing the beginning time and the service time. Thus, we can calculate the overall estimated latency.

In the feedback scheme, we formulate the arrival rate of workloads, the server's serving rate and the user's abandon rate before using the Markov chain to model the two queues. After calculating queue length and the expected waiting time, we equalize workload latencies by adjusting the resource allocation. The detailed adjustment description will be introduced in the next subsection.

Overall, D^{3}G converts the performance optimization problem into a resource allocation problem. By estimating latencies of different network services, D^{3}G mitigates the imbalance by adjusting the allocated resources. The real time estimation algorithm and the intelligent feedback scheme make D^{3}G work efficiently and automatically.

Adjusting Resource Allocation

To calculate the resources allocated to interactive workloads and non-interactive workloads, we model the queuing problem in the adjusting scheme. To analyze the arrival and leaving rates of requests, we adopt a Markov model to represent state transmissions of the two queues. By modeling different distributions on service time and abandon rate, we can calculate the latency expectation of various workloads. Finally, the feedback scheme could adjust resource allocation to ensure fairness.

The arrival process of requests is a discrete-time random process, and the number of requests in the future is only related to the number at present, i.e., the queue can be formulated by a Markov chain, and the queue length can be calculated. For any specific server, when a request arrives, the queue length increases by 1; when a request abandons the queue or a service is finished, the queue length decreases by 1. Thus, with the arrival rate, service time and abandoning rate, we can model the queues in any particular servers as M/M/1 queues [14] and determine the queue length. As described before, the service time is in an exponent distribution, and each service is mutually independent from each other, so the waiting time of a request is a convolution. So far, the expectation of waiting time on one server can be calculated.

Recall that interactive requests will be pending in queues for multiple times in different servers, and non-interactive requests only need to be pending once. This may lead to intolerable latency for delay-sensitive workloads. We solve this problem by adjusting resource allocations. With the expected waiting time, we assume the transfer time on a server is in a Gaussian distribution [15], and then make the overall latency of the two queues equal to each other. Thus, we can work out the allocation rates to the interactive workloads and non-interactive workloads. With this adjusted allocation, interactive workloads from delay-sensitive applications can enjoy reduced latency that is within user tolerance.

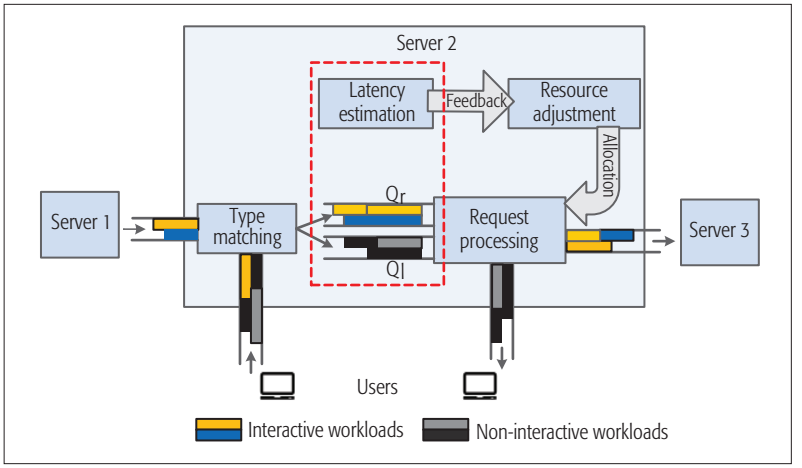

FIGURE 3. The framework of the D3G scheme.

Deployment

We implement D^{3}G in the servers of Baidu networks, and the algorithm is written in C language. The servers use the Linux operating system and are configured with tomcat webservers based on Java. We choose four servers that are configured with 4 GB of memory, two cores, and 100 Mb/s public network bandwidth. As to the clients, there are 36 end-hosts and each is configured with an Intel i5 1.7 GHz CPU and 2 GB of memory. All these end-hosts are constantly sending either interactive or non-interactive requests to those servers. The interactive workloads need to be served in each server, and the non-interactive workloads can be processed by only one server. We conduct a series of experiments in the next section:

- Overall performance: we conduct a series of experiments, measuring the average response time of both interactive and non-interactive workloads under D^{3}G versus the state-of-the-art scheme without D^{3}G.
- Algorithm dynamism: we test the algorithm's performance under a dynamic scenario.
- System scalability: we evaluate the optimization of D^{3}G under expanding scales.

Evaluation

As described in the deployment section, we conduct three groups of experiments to test the algorithm's efficiency and evaluate average response time and service performance under different network environments. Recall the example given in a previous section that the interactive workloads are actually suffering from seven times longer latency compared to non-interactive workloads.

The experiment results in this section show that D^{3}G significantly reduces latency for time-sensitive workloads. At the same time, non-interactive workloads are not affected seriously and are still enjoying shorter latencies. Besides verifying the effectiveness of D^{3}G algorithm, we also prove the potential practicability for large-scale deployment.

Overall Performance

In this subsection, we design several groups of experiments to evaluate the performance of D^{3}G under different network environments. We start the experiment in a 9-to-1 model. In this case,

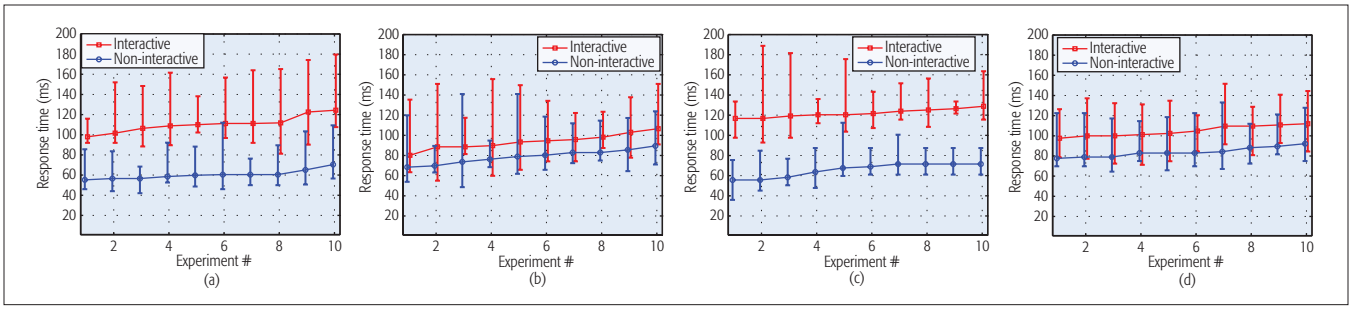

FIGURE 4. Average response time of both interactive and non-interactive workloads: a) latency without D3G when requests are (100KB, 200KB]; b) latency with D3G when requests are (100KB, 200KB]; c) latency without D3G when requests are 200KB and above; d) latency with D3G when requests are 200KB and above.

every nine end-hosts keep sending interactive and non-interactive requests to one server. We record the response time of different workloads under different scenarios.

For small requests, we set the workload length from 100 *KB* to 200 *KB*, conduct the experiments 200 times, and calculate the average response time per 20 experiments with upper and lower error bars. Figure 4a shows the response time of both interactive and non-interactive requests before deploying D^{3}G; Fig. 4b shows the response time after deploying D^{3}G. These figures show that D^{3}G can significantly reduce the response time of interactive workloads (from 110 ms to 95 ms on average), and the latency for non-interactive workloads is not seriously affected (from 60 ms to 75 ms).

For large requests, we set the workload length to at least 200 *KB*. Figure 4c shows the results without D^{3}G, from which we can observe that the average latency for interactive workloads is about 120 ms and that for non-interactive workloads is about 75 ms, indicating that interactive requests are suffering from 1.6 times longer latency than non-interactive requests. Figure 4d shows the optimized results after deploying D^{3}G, where interactive workload latency is reduced by 33 percent (to 80 ms) on average with minimal impact on non-interactive workload latency.

From these results, we can conclude that D^{3}G works well in accelerating interactive workloads under various circumstances.

Algorithm Dynamism

To evaluate our algorithm in dynamic scenarios, we simulate a dynamic situation to verify the real-time efficiency of D^{3}G.

We send only non-interactive requests in the previous 60 s, and then begin to send interactive requests at the 60th s and stop sending at the 130th s. Figure 5a shows the average delay of this dynamic process. Latency is quite long for a short period of time (from 60 s to 70 s). Then it begins to drop because more resources are allocated to the interactive queue. When interactive workloads stop at the 130th s, non-interactive workload latency drops.

From these experiments, we can conclude that D^{3}G accelerates interactive workloads while not seriously affecting non-interactive workloads.

System Scalability

Finally, we extend the experiment scales and increase concurrency to test the algorithm's scalability. We speed up the request sending rate, and Fig. 5b shows the average latency on various scales. When there are 50 concurrent requests, the average latency is about 65 ms for non-interactive workloads and 80 ms for interactive ones. When the number of concurrent requests increases to 500, the average latencies are about 150 ms and 170 ms, respectively. These results indicate that our algorithm is extensible in large-scale systems. Furthermore, if the interactive workloads are handled by more cloud servers, the latency without D^{3}G will become even higher (as shown in the case in the previous section), and our algorithm optimization will be more obvious.

From the above deployment and evaluations, we can conclude that D^{3}G successfully reduces the latency of interactive workloads to a reasonable range with no distinct impact on non-interactive workloads, even in expanding scales. We believe that the main idea of D^{3}G, to reduce latency of interactive workloads from time-sensitive applications, will soon be adopted by current microservices.

Conclusion

For cloud-based service chains, we measure and analyze their performance in Baidu networks, and the results show that these delay-sensitive microservice-like applications are suffering from long latency due to the extra delay from the multiple stand-alone components.

In this article, we propose a new algorithm called Dynamic Differentiated service for Delay-Guarantee (D^{3}G), which aims at reducing the overall latency for chained applications while ensuring workload fairness. To this end, we design two queues in servers. One is for interactive requests, and the other is for non-interactive requests. To make the latency within user tolerance, the latency estimation algorithm pre-calculates interaction latency. Furthermore, to guarantee fairness, we introduce a feedback control scheme based on resource allocation to ensure the performance of non-interactive workloads. A wide range of detailed evaluation results demonstrate that D^{3}G succeeds in accelerating chained services and ensuring workload fairness. As microservice-like applications have many obvious advantages such as clean boundaries, better system scalability and lower operating costs, they are attracting increasing attention. We believe that D^{3}G will further reveal its effectiveness along with the development of service chains.

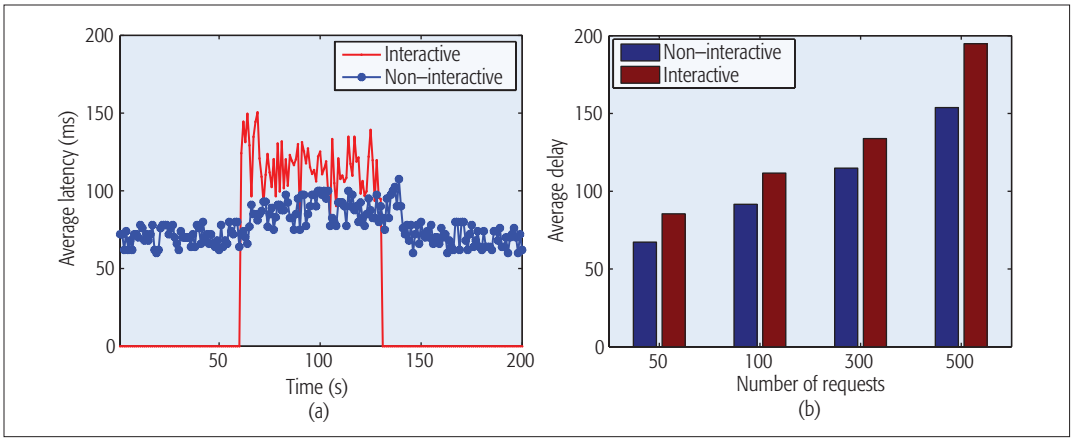

FIGURE 5. Performance for different parameters: a) average delay for dynamic scenario; b) average delay for different scales.

Acknowledgment

This work was supported by the National Natural Foundation of China (61472212 and 61572278), EU Marie Curie Actions CROWN (FP7-PEOPLE-2013-IRSES-610524), the National Key R&D Program of China (2016YFB0800102), and the R&D Program of Shenzhen (JCYJ20170307153259323).

References

[1] Y. Guo *et al.*, "Energy and Network Aware Workload Management for Sustainable Data Centers with Thermal Storage," *IEEE Trans. Parallel and Distributed Systems*, 2014, vol. 25, no. 8, pp. 2030–42.
[2] M. Alizadeh *et al.*, "pfabric: Minimal Near-Optimal Datacenter Transport," *ACM SIGCOMM Computer Communication Review*, ACM, 2013, vol. 43, no. 4, pp. 435–46.
[3] F. R. Dogar *et al.*, "Decentralized Task-Aware Scheduling for Data Center Networks," *ACM SIGCOMM Computer Communication Review*, ACM, 2014, vol. 44, no. 4, pp. 431–42.
[4] Y. Zhang *et al.*, "Towards Shorter Task Completion Time in Datacenter Networks," *Proc. 34th IEEE Int'l. Conf. Computing and Communications*, IEEE, 2015, pp. 1–8.
[5] M. Mauve *et al.*, "Local-Lag and Timewarp: Providing Consistency for Replicated Continuous Applications," *IEEE Trans. Multimedia*, 2004, vol. 6, no. 1, pp. 47–57.
[6] S. D. Webb, S. Soh, and W. Lau, "Enhanced Mirrored Servers for Network Games," *Proc. 6th ACM SIGCOMM workshop on Network and System Support for Games*, ACM, 2007, pp. 117–22.
[7] K. H. Vik, P. Halvorsen, and C. Griwodz, "Multicast Tree Diameter for Dynamic Distributed Interactive Applications, *Proc. 27th IEEE Conf. Computer Commun., INFOCOM 2008*, IEEE, 2008, pp. 1597–1605.
[8] J. Guo *et al.*, "A Cooperative Game Based Allocation for Sharing Data Center Networks, *Proc. IEEE 2013 INFOCOM*, IEEE, 2013, pp. 2139–47.
[9] K. Xu *et al.*, "Online Combinatorial Double Auction for Mobile Cloud Computing Markets," *Proc. 2017 IEEE Int'l. Performance Computing and Communications Conf. (IPCCC)*, IEEE, 2014, pp. 1–8.
[10] Y. Zaki *et al.*, "Dissecting Web Latency in Ghana," *Proc. 2014 Conf. Internet Measurement*, ACM, 2014, pp. 241–48.
[11] E. Pujol *et al.*, "Back-Office Web Traffic on the internet," *Proc. 2014 Conf. Internet Measurement*, ACM, 2014, pp. 257–70.
[12] H. Wang *et al.*, "On Design and Performance of Cloud-Based Distributed Interactive Applications," *Proc. 2014 22nd IEEE Int'l. Conf. Network Protocols (ICNP)*, IEEE, 2014, pp. 37–46.
[13] J. Carlstrom and R. Rom, "Application-Aware Admission Control and Scheduling In Web Servers," *Proc. INFOCOM 2002, Twenty-First Annual Joint Conf. IEEE Computer and Communications Societies*, IEEE, 2002, 2, pp. 506–515.
[14] M. Mitzenmache, "The Power of Two Choices in Randomized Load Balancing," *IEEE Trans. Parallel and Distributed Systems*, 2001, vol. 12, no. 10, pp. 1094–1104.
[15] E. Pebesma *et al.*, "INTAMAP: The Design and Implementation of an Interoperable Automated Interpolation Web Service," *Computers & Geosciences*, 2011, vol. 37, no. 3, pp. 343–52.

Biographies

Yuchao Zhang received the Bachelor of Science degree in computer science and technology from Jilin University, China in 2012., and her Ph.D. degree from the Department of Computer Science & Technology of Tsinghua University, Beijing, China in 2017. Currently she works at Beijing University of Posts and Telecommunications. Her research interests include cloud computing, large-scale datacenter networks, high-speed networks and network function virtualization.

Ke Xu [M'02, SM'09] received his Ph.D. from the Department of Computer Science & Technology of Tsinghua University, Beijing, China, where he serves as a full professor. He has published more than 100 technical papers and holds 20 patents in the research areas of next generation Internet, P2P systems, Internet of Things (IoT), and network virtualization and optimization. He is a member of ACM and has guest-edited several special issues in IEEE and Springer journals. Currently, he is holding a visiting professor position at the University of Essex.

Haiyang Wang is an assistant professor in the Department of Computer Science at the University of Minnesota Duluth, USA. His research interests include cloud computing, peer-to-peer networking, social networking, big data and multimedia communications.

Qi Li [M'12] received the B.Sc. and Ph.D. degrees in computer science from Tsinghua University, Beijing, China, in 2003 and 2012, respectively. His research interests include network architecture, protocol design, and system and network security.

Tong Li received his B.S. degree from the Department of Computer Science of Wuhan University, Hubei, China in 2012, and his Ph.D. degree from the Department of Computer Science & Technology of Tsinghua University in 2017. Currently he works at Huawei Company. His research interests include network virtualization and resource management, network science, and P2P systems.

Xuan Cao received his B.S. and Master degrees from the Department of Computer Science of Nanjing University of Science and Technology in 2008 and 2011, respectively. He has been with Beijing Baidu Netcom Science Technology Co., Ltd. since then. His research interests include resource management, network virtualization, pattern recognition and intelligent systems.