Unfair Data Centers for Fun and Profit

Akshitha Sriraman

University of Michigan
akshitha@umich.edu

1. INTRODUCTION

On-Line Data Intensive (OLDI) applications, such as web search, advertising, online retail, social networking, and software-as-a-service form a major fraction of data center applications [1, 2]. OLDI applications must often meet stringent tail ($99^{th} + \%$) latency constraints expressed in the form of soft real-time deadlines called Service Level Objectives (SLOs) [3, 4, 5, 6, 7]. For instance, web search operators seek an SLO that is less than 300 ms, so that search results feel instantaneous to end users [8, 9].

One significant source of OLDI latency tails is queuing delays: network interference [10], congestion, hostile co-runners [11], OS scheduler inefficiencies [3], etc. can cause request queuing that precipitate SLO violations. Even if as few as one machine in 10,000 straggles due to queuing delays, up to 18% of requests can experience a high latency [10].

Prior works [10, 12] have aimed to reduce queuing delays by introducing explicit prioritization of requests to OLDI applications. However, these prior works as well as modern data centers operate under the assumption that (1) all end-users have the same stringent SLO expectation and (2) meeting these user-agnostic fixed SLOs determines the Quality of end-user Experience [13, 14, 15, 16] (QoE). In this paper, we argue that whereas stringent SLOs must be met for some end-users who expect to perceive an "instantaneous" response, there may exist many end-users who have a more tolerant SLO requirement i.e., QoE is a subjective metric. Relaxing the SLO constraint of a more tolerant end-user's request in real-time can reduce queuing delays by facilitating better request prioritization, thereby improving the performance and energy efficiency of a data center.

We propose investigating user traits and their correlation (if any) with acceptable OLDI response wait times. For example, does age, nature of employment, geographical location, sex, race, or political orientation (or a combination of subsets of these features) affect a user's SLO tolerance? A common opinion is that an elderly user may be more tolerant [17] than a younger one. Perhaps such elderly users may be willing to wait for longer than 300 ms for a web search response than a younger user. In this paper, we propose prioritizing requests in an OLDI service's request queue based on the SLO tolerance threshold of the end-user who sent the request.

An "unfair" data center that prioritizes requests in a user-cognizant manner might be socially controversial. For example, several modern online retail, online advertisement, and social media platforms that base their revenue on tailoring service content in a user-cognizant manner [18] have been subject to legal scrutiny. However, we argue that the system proposed in this paper is not inherently socially controversial or offensive; the *user traits* considered might raise ethical concerns. Nevertheless, determining ethical user traits and their SLO implications is still an interesting scientific question.

In this paper, we propose investigating (1) which traits, if any, are correlated to a user's data center performance expectations? and (2) of these traits, which are ethically acceptable for use in an "unfair" data center? Given a set of ethically-acceptable traits, we propose developing an "unfair" data center, a data center scheduler that sets request-specific SLOs and routes requests based on the *SLO tolerance threshold* of the end-user who sent the request. While a user's tolerance threshold may seem like a qualitative metric, we aim to quantify tolerance by considering users' service abandonment rates [19]. Computing user-specific abandonment rates is relatively simple in applications that establish the identity of the end-user during login (e.g., social media, online retail, online advertisement, email, etc).

The "unfair" data center scheduler may also be extended to set user-specific SLOs based on secondary factors (e.g., time of day, day of the week, web pages to be displayed, mobile vs. desktop usage, dedicated mobile apps vs. mobile browser usage, etc). For instance, a user might anticipate a faster response when they are at work during the day than during times of respite at night.

2. MOTIVATION

Most modern OLDI applications must meet fixed response latency SLO constraints. For example, web search operators categorize response latencies greater than 300 ms as SLO violations. In contrast, we propose setting user-tolerance-threshold—cognizant service SLOs, to reduce queuing delays and improve data centers' perf/watt budgets.

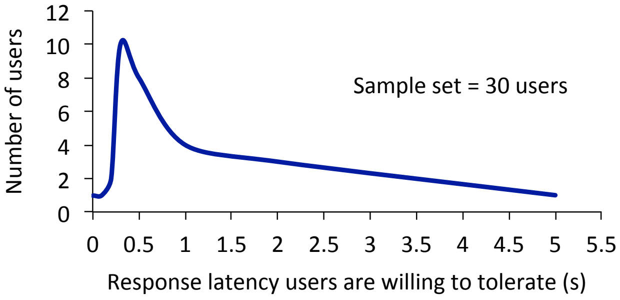

Figure 1: CDF of the response latency users are willing to tolerate (sample set size of 30 users): several users are willing to tolerate a response latency greater than the typical SLO of 300 ms

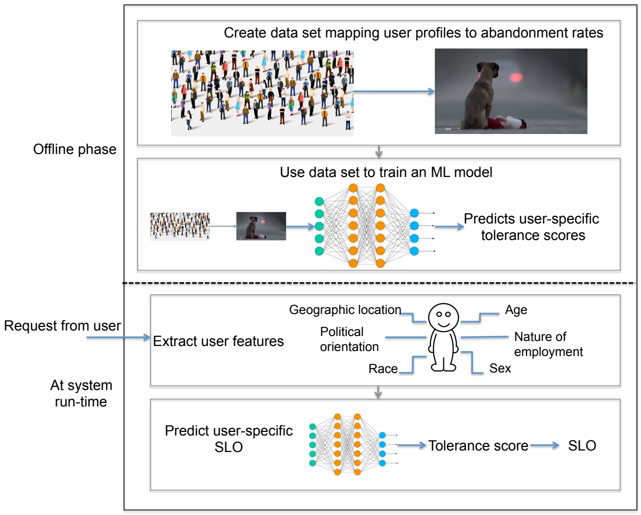

Figure 2: The Unfair Data Center scheduler design

The idea of an "unfair" data center is based on the hypothesis that different people may have different response latency tolerance thresholds. To verify this hypothesis, we conducted a preliminary study with thirty individuals from relatively diverse backgrounds. Each individual was asked to enter the time they were willing to wait for a web search response, in a user poll. The results of the poll are shown in Fig. 1. From the graph, we see that setting flexible user-specific SLOs can provide a latency head room of $\sim$ 60%, thereby allowing the more latency-critical requests to drain faster from request queues. This increase in latency head room directly translates to better (1) application throughput and (2) data center performance and energy efficiency [4].

3. SYSTEM OVERVIEW

In this section, we describe the "unfair" data center's design and its implications.

3.1 System Design

We describe the "unfair" data center scheduler's design for OLDI services where user-specific data is available after login. For OLDI applications that are ignorant of user information (e.g., web search), the "unfair" data center's design can be extended to build machine learning models that estimate user profiles based on IP addresses associated with requests.

The "unfair" data center scheduler's operation comprises two phases: an offline training phase and an online SLO prediction phase, as depicted in Fig. 2. During the offline training phase, the scheduler uses run-time statistics trackers [20, 21, 22, 23] available in modern data centers to assimilate a data set relating a user and their tolerance level-influencing traits (may include nature of mobile phone, age, race, sex, geographical location, time of day, day of the week, mobile vs. desktop usage, etc) to their abandonment rate. The scheduler subsequently uses this data set to train a machine learning model offline.

During system run-time, the scheduler tracks every request sent to the OLDI application. When a request arrives, the scheduler first extracts user-specific information that can impact service abandonment (e.g., age) based on secondary features (e.g., time of day) monitored by fleet-wide profiling tools. We expect this extraction step to not incur any additional overhead as several OLDI applications already perform user profile extractions to display user-specific ads, content [18], etc. The scheduler then feeds the extracted user profile to the machine learning model to infer the user's predicted tolerance score. The scheduler sets an SLO for the user's request based on the predicted tolerance score.

3.2 System Design Implications

The "unfair" scheduler may induce some SLO violations when the predicted SLO latency is greater than the actual user SLO tolerance threshold. We propose mitigating such incorrect predictions by tuning the ML model at frequent intervals. Fleet-wide profiling tools [20] can be used to track SLO violations. Consequently, user profiles corresponding to requests that caused SLO violations can be used to re-train the ML model.

4. RELATED WORK

Prior works have studied the impact of user behavior on abandonment rates. Sitaraman et al. [24] study the impact of stream quality on user behavior in video streaming services. They find that: (1) users abandon a video if the startup time is more than two seconds, (2) every incremental delay of 1 second increases the abandon rate by 5.8%, (3) users with better internet connection have less patience for startup delays, and (4) users on mobile devices have the most patience. Brutlag et al. [25] compare two mock search engines that only differ in branding and response latency. Their study shows that when the response time is over 3 seconds, the user is 1.5 times more likely to choose the faster search engine. However, these prior works relate user interactions to abandonment rates instead of relating user-specific profile characteristics to abandonment rates and SLOs.

5. CONCLUDING RESEARCH QUESTIONS

"Unfair" data centers raise several interesting research questions that require detailed exploration to improve modern data centers' performance and energy efficiency.

(1) Is the underlying assumption that *all* end-users expect extremely stringent latency constraints inherently true?

(2) Which user traits, if any, are correlated to a user's response latency tolerance threshold?

(3) Of these traits, which are ethically acceptable for use in an "unfair" data center?

(4) To what effect can incorrect user-specific SLO predictions degrade OLDI tail latency?

(5) What are the latency overheads associated with tracking secondary factors such as time of day?

Furthermore, the key idea of user-specific computation can be applicable beyond data center SLOs. For example, a processor's Dynamic Voltage and Frequency Scaling (DVFS) settings can be adjusted based on the performance requirement of the end-user to reduce processor power consumption [26]. Similarly, wrist-band sensors can be designed to predict stress based on the end-user [27].

6. ACKNOWLEDGEMENT

We thank the anonymous reviewers for their valuable feedback. We acknowledge Thomas F. Wenisch, Amrit Gopal, Kevin Loughlin, Brandon West, and Amritha Varshini for their insightful suggestions that helped improve this work.

7. REFERENCES

[1] D. Meisner, C. M. Sadler, L. A. Barroso, W.-D. Weber, and T. F. Wenisch, "Power Management of Online Data-intensive Services," in *International Symposium on Computer Architecture*, 2011.

[2] A. Sriraman, A. Dhanotia, and T. F. Wenisch, "SoftSKU: Optimizing Server Architectures for Microservice Diversity @Scale," in *The International Symposium on Computer Architecture (to appear)*, 2019.

[3] A. Sriraman and T. F. Wenisch, "μSuite: A Benchmark Suite for Microservices," in *IEEE International Symposium on Workload Characterization*, 2018.

[4] A. Sriraman and T. F. Wenisch, "μTune: Auto-Tuned Threading for OLDI Microservices," in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, 2018.

[5] A. Mirhosseini, A. Sriraman, and T. F. Wenisch, "Enhancing server efficiency in the face of killer microseconds," in *IEEE International Symposium on High Performance Computer Architecture*, 2019.

[6] A. Mirhosseini, A. Sriraman, and T. F. Wenisch, "Hiding the Microsecond-Scale Latency of Storage-Class Memories with Duplexity," in *Proceedings of the 10th Annual Non-Volative Memories Workshop*, 2019.

[7] A. Sriraman and T. F. Wenisch, "Performance-Efficient Notification Paradigms for Disaggregated OLDI Microservices," in *Workshop on Resource Disaggregation*, 2019.

[8] B. Vamanan, J. Hasan, and T. Vijaykumar, "Deadline-aware Datacenter TCP (D2TCP)," in *ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2012.

[9] A. Sriraman, S. Liu, S. Gunbay, S. Su, and T. F. Wenisch, "Deconstructing the Tail at Scale Effect Across Network Protocols," *The Annual Workshop on Duplicating, Deconstructing, and Debunking*, 2016.

[10] M. P. Grosvenor, M. Schwarzkopf, I. Gog, R. N. M. Watson, A. W. Moore, S. Hand, and J. Crowcroft, "Queues don't matter when you can jump them!," in *Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation*, 2015.

[11] J. Mars, L. Tang, R. Hundt, K. Skadron, and M. L. Soffa, "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations," in *Proceedings of the 44th annual IEEE/ACM International Symposium on Microarchitecture*, 2011.

[12] S. Blagodurov, D. Gmach, M. Arlitt, Y. Chen, C. Hyser, and A. Fedorova, "Maximizing server utilization while meeting critical SLAs via weight-based collocation management," in *IFIP/IEEE International Symposium on Integrated Network Management*, 2013.

[13] I. Arapakis, X. Bai, and B. B. Cambazoglu, "Impact of Response Latency on User Behavior in Web Search," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014.

[14] A. Bouch, N. Bhatti, and A. Kuchinsky, "Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service," in *ACM Conference on Human Factors and Computing Systems*, 2000.

[15] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers Not to the Hippo," in *International Conference on Knowledge Discovery and Data Mining*, 2007.

[16] "Latency is everywhere and it costs you sales - how to crush it." http://highscalability.com/blog/2009/7/25/ latency-iseverywhere-and-it-costs-you-sales-how-to-crush-it.html.

[17] "How do you think elderly people can become more tolerant and open-minded?." https://www.quora.com/How-do-you-think-elderly-people-can-become-more-tolerant-and-open-minded.

[18] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. n. Candela, "Practical lessons from predicting clicks on ads at facebook," in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 2014.

[19] H. Nam, H. Schulzrinne, H. Nam, K.-H. Kim, H. Schulzrinne, M. Varela, H. Nam, H. Schulzrinne, T. Mäki, H. Nam, *et al.*, "Youslow: What influences user abandonment behavior for internet video?," *Tech. report*, 2017.

[20] D. Borthakur, J. Gray, J. S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, *et al.*, "Apache Hadoop goes realtime at Facebook," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 2011.

[21] T. Pelkonen, S. Franklin, J. Teller, P. Cavallaro, Q. Huang, J. Meza, and K. Veeraraghavan, "Gorilla: A fast, scalable, in-memory time series database," *Proceedings of the VLDB Endowment*, 2015.

[22] A. S. Aiyer, M. Bautin, G. J. Chen, P. Damania, P. Khemani, K. Muthukkaruppan, K. Ranganathan, N. Spiegelberg, L. Tang, and M. Vaidya, "Storage infrastructure behind Facebook messages: Using HBase at scale," *IEEE Data Eng. Bull.*, 2012.

[23] G. Ren, E. Tune, T. Moseley, Y. Shi, S. Rus, and R. Hundt, "Google-wide profiling: A continuous profiling infrastructure for data centers," *IEEE micro*, 2010.

[24] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," in *Proceedings of the Internet Measurement Conference*, 2012.

[25] J. D. Brutlag and M. H. Stone, "User preference and search engine latency," in *JSM Proceedings: Quality and Productivity Research Section*, 2008.

[26] L. Yang, R. P. Dick, G. Memik, and P. Dinda, "HAPPE: Human and Application-Driven Frequency Scaling for Processor Power Efficiency," *IEEE Transactions on Mobile Computing*, 2013.

[27] B. Egilmez, E. Poyraz, W. Zhou, G. Memik, P. Dinda, and N. Alshurafa