Impact of Delayed Caching on Hit-ratio of ICN Router

Feri Fahrianto[1,2] and Noriaki Kamiyama[3]

[1]Graduate School of Engineering, Fukuoka University, Japan
[2]Faculty of Science and Technology, Syarif Hidayatullah Jakarta State Islamic University, Indonesia
[3]College of Information Science and Engineering, Ritsumeikan University, Japan

1 Introduction

A caching system is implemented in the Information-Centric Networking (ICN) router to improve network connectivity. The Content Store (CS) is a component that allows users to manage temporary data packets from a certain producer. The desired data content necessitates the consumer to send an initial interest packet invocation as a request from a client. Apparently, the hit-ratio in CS is affected by the average arrival time of interest packets and producer response time, besides the CS capacity size. When the average request time from the consumer is faster than the producer's response time, it will cause a condition called delayed caching. Therefore, the paper investigates the consequences of delayed caching in the ICN router, particularly in terms of hit-ratio reduction in the CS. The different values of request skewness (α) representing the request content popularity are simulated against the arrival request time and producer response time in the simulation testbed. Moreover, the testbed result is compared with the analytical hit-ratio estimation model.

2 Delayed caching

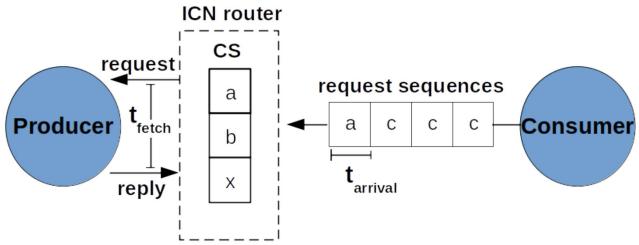
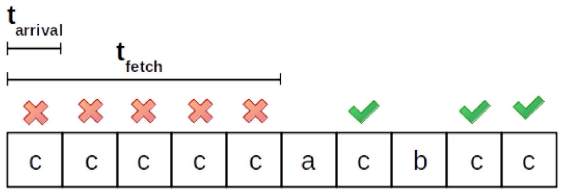


Figure 1: Caching with delay in Content Store (CS)

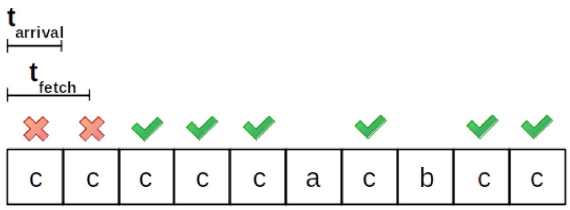
Figure 1 depicts a typical cache with a delay that occurred in the ICN router. When a request for data content arrives, it is unavailable in the CS. The ICN router sends a request to a producer to get the data object. The retrieval takes some non-zero amount of time known as the producer response time, or t_{fetch}, and the average arrival time between consumer interest packets, known as $t_{arrival}$. If a new request for the same object arrives at the CS before t_{fetch} elapses, this new request suffers from a cache miss in the CS. If the condition repeatedly occurs in the ICN router, it causes a significant underperformance of the CS hit ratio.

3 Hit-ratio degradation

The hit ratio is defined as the number of requests satisfied by the CS divided by the number of all requests arrived at the CS. The higher the hit ratio is or the lower the miss ratio is, the more requests are satisfied by the CS.



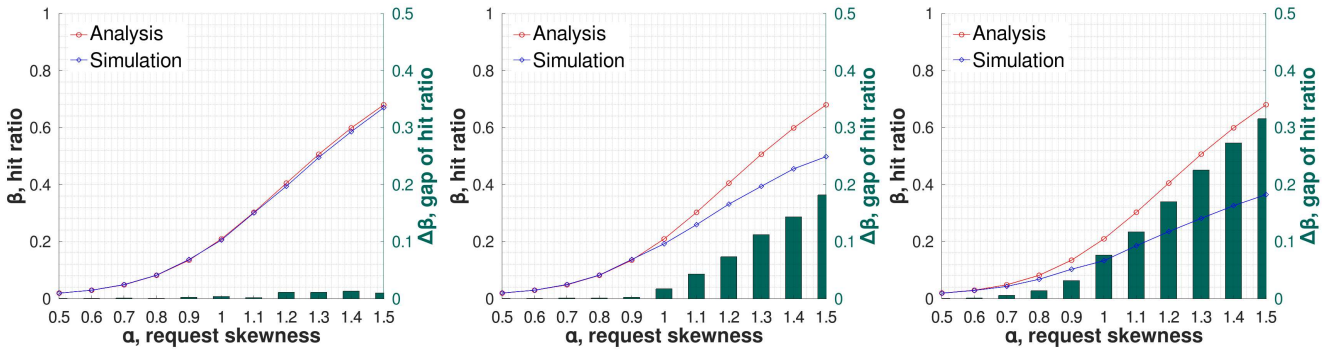
(a) Cache miss in $t_{fetch} \simeq 5 \times t_{arrival}$



(b) Cache miss in $t_{fetch} \simeq 1.5 \times t_{arrival}$

Figure 2: The hit-ratio degradation in different length of t_{fetch} relative to $t_{arrival}$

The hit-ratio degradation that occurs as a result of caching delay is depicted in Figure 2. Figure 2(a) shows the case when t_{fetch} is 5 times of $t_{arrival}$, whereas Figure 2(b) shows the case when t_{fetch} is 1.5 times of $t_{arrival}$. In the first case, just three requests for object c are satisfied by the CS, whereas six requests for object c are satisfied by the CS in the second case. The longer t_{fetch} relative to $t_{arrival}$ can cause more cache miss. As a result, the hit-ratio degradation occurs in the CS of the ICN router. Since CS in the ICN router implements a specific replacement algorithm such as Least Recently Used (LRU), we used the approximation of hit ratio introduced by using Che's approximation [2] to use as a reference measurement in our simulation testbed. The hit ratio for particular content i, $\beta(i)$,



(a) In case of $t_{fetch} \simeq t_{arrival}$ (b) In case of $t_{fetch} \simeq 10 \times t_{arrival}$ (c) In case of $t_{fetch} \simeq 100 \times t_{arrival}$

Figure 3: The comparison of hit-ratio (β) between analytical model and simulation testbed in three different value of t_{fetch} relative to $t_{arrival}$.

could be estimated by

$$\beta(i) = 1 - e^{-q(i)tc} \tag{1}$$

The symbol $q(i)$ is the ratio of requests for content i, and t_c is the unique root of the equation

$$C = \sum_{i=1}^{M}(1 - e^{-q(i)tc}) \tag{2}$$

where C is the storage capacity of CS, and M is the number of unique content items, i.e., content catalog-size. So the total hit-ratio, β, can be expressed by

$$\beta = \sum_{i=1}^{M} q(i)\beta(i). \tag{3}$$

4 Numerical evaluation

An experiment using a simulator testbed was conducted to measure the hit-ratio reduction at CS with the varying value of t_{fetch} relative to $t_{arrival}$. Table 1 shows the major hardware and software parameters used in the experiment.

Table 1: Setting values of main parameters

Paramater	Value
Number of content items	1000
Cache size	10
Request skewness (α)	0.5 - 1.5
t_{fetch}	10 μs, 100 μs, 1000 μs
$t_{arrival}$	10 μs
CPU	AMD Ryzen 9 3900X
OS	Ubuntu 20.04 LTS
Programming language	Python 3.8

We obtained the data from 100000 requests sent by the consumer. β, the total hit-ratio and, $\Delta\beta$, the gap between the hit ratio obtained by the analytical model and that obtained by the simulator, are plotted against the request skewness, α, in Figure 3 for each of the three values of t_{fetch}. In general, the result shows that the smaller α creates smaller gap compare to higher α in every different case of t_{fetch} relative to $t_{arrival}$. Furthermore, $\Delta\beta$ monotonically increased as t_{fetch} increased. The average of $\Delta\beta$ were about 1%, 15% and 30% when $t_{fetch} \simeq t_{arrival}$, $t_{fetch} \simeq 5 \times t_{arrival}$, and $t_{fetch} \simeq 5 \times t_{arrival}$ respectively. When more requests concentrated on a limited set of popular content items, the cache hit ratio was more sensitive to delayed caching. This is because that more requests for popular content are more likely to be requested during the t_{fetch} interval and being suffered from cache miss in the CS.

5 Conclusion

The hit-ratio degradation in the ICN router can occur because the response time from the producer is longer than the incoming request time from the consumer. Thus, the more popular the content requested from the consumer, the higher the risk of experiencing a hit-ratio degradation since the popular content is more likely to be requested within the producer's response time interval. In future, we will investigate a method of suppressing the hit-ratio degradation caused by delayed caching in the CS of ICN routers.

Acknowledgement This work was supported by JSPS KAKENHI Grant Number 18K11283 and 21H03437, and the Ministry of Religious Affairs of the Republic of Indonesia (MORA 5000 Doktor).

References

[1] Andrea Detti, Lorenzo Bracciale, Pierpaolo Loreti, Nicola Blefari Melazzi, "Modeling LRU Cache with Invalidation," Comput. Netw. 134, C (April 2018), Pages 55–65, 2018.

[2] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web Caching Systems:Modeling, Design and Experimental Results," IEEE J. Selected Areas of Commun., vol.20, no.7, pp.1305-1314, Sep. 2002.

[3] N. Atre, J. Sherry, W. Wang, and D. S. Berger,"Caching with Delayed Hits," In Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM '20). Association for Computing Machinery, New York, NY, USA, 2020.

[4] S. Selvakumar, Swarup Kumar Sahoo, Vijay Venkatasubramani, "Delay Sensitive Least Frequently Used Algorithm for Replacement in Web Caches," Computer Communications, Volume 27, Issue 3, Pages 322-326, 2004.