EXPERIMENTAL RESEARCH OF SERVICE SYSTEM

Liudvikas Kaklauskas
Šiauliai State University off Applied Sciences
Lithuania

Annotation

Most of the modern service systems have different performance channels. Most of them are ineffective because queries are distributed to channels in random order. In order to use them efficiently it is necessary to choose the right strategy for query distribution. In this article we take experimental research with service system having two different performance channels. The study found, that service system efficiency belongs from flow intensity and from detention buffer length.

Key words: service system, different performance channels, buffer, query, heterogeneous network.

Introduction

Service systems are widely used in business and technology. Efficient service systems are necessary in order to save resources as well as to serve as much clients as possible. Currently there are many various service strategies which have their own pros and cons. The known service strategies are suited for systems with equal performance channels, however, in practice service systems have different performance channels (Erlong, 1909, Sakalauskas, 2000, Iversen, 2011). For instance, at the shop every cashier's efficiency depends on her experience in this field, therefore her service speed can be several times higher than the new cashier who is working for several days. The similar situation is encountered in computers where work is done by CPU and GPU. Thus their capacity may differ up to 100 times, therefore it is necessary to allocate them the work optimally (Bilel, etc., 2012, Hetherington, etc., 2012).

However the systems with different performance channels are not discussed sufficiently in scientific literature as in this case the received mathematical patterns are greatly difficult (Efrosinin, Rykov, 2008, Rykov, Efrosinin, 2009). The well-known works by E. R. Larsen (Sankaranarayanan, Larsen, van Ackere, Delgado, 2010) as well as Rybinovitch (Rubinovitch, 1985, Mokaddis, Matta, El Genaidy, 1998) discussed the service systems with two different performance channels, however not many results are provided and without clear grounding.

When the capacities of processors vary with little difference, then the service disciplinary might be the first to come the first to be served, and the query goes straight to the free channel or waits in line until another channel clears up. It is noted that under high capacity ratio this disciplinary slows the system work (Rykov, Efrosinin, 2008). In case of high ratio, it appears that it is useful to install detention buffer as sometimes it is more efficient to wait until the high speed channel gets free eventhough the inefficient channel was free. Systems with detention buffers were discussed in works, as well as some of the trials have been received in Rubinovicius's work. Recently the relevance of research of these systems has been increased by creating multiprocessor calculation systems, combining CPU (A Central Processing Unit) and GPU (A Graphics Processing Unit) processors (Hetherington, & etc., 2012, Kadjo, & etc., 2015), by creating combined service systems of high and low capacity networks (Rykov, Efrosinin, 2004, Yue, &etc., 2009, Efrosinin, Sztrik, 2011). As their capacity may vary up to 100 times therefore there is an actual task to discuss systems in which the ratio of capacity of used processors is increasing and we can apply respective asymptotic extensions. In the scientific literature about the systems with two different performance channels mostly analyzes theoretical side of problem, but no used experimental research by simulation. This work it is discussed experimental research of service system having two different performance channels.

1. Two-channel service systems characteristics

A discussed service system consists of two different performance channels as well as one detention buffer. Assuming that the time length between two adjacent queries is falling within the discussed service systems, allocated under the Puason's Law with parameter λ and the length of service is allocated also under this Law with parameters μ_1 and μ_2 (faster channel and slower channel). Assuming that the queries are served in sequence, i.e. first came – first served (Gelenbe, Pujjole, 1999, Xiaolong, Geyong, 2009).

If the query after being released into the system finds a free performance channel, it is served immediately; otherwise it goes to detention buffer of *k* length where it waits until the efficient channels gets free. One query is served by one channel. If all places in detention buffer are occupied, the query rejects the service of performance channel and transfers to the slow channel. If the slow channel is occupied as well, the application waits in m row of finite length. If all places are occupied, the query rejects and not served. This strategy is suitable only when the coefficient of flow volume $\rho = \frac{\lambda}{\mu_1+\mu_2}$ is lower than 1, because in case of higher intensity it is not efficient to keep free channels even though they are slow.

In the service system with two different performance channels we use denotes and formulas discussed in L. Kaklauskas, L. Sakalauskas and V. Denisovas article (Kaklauskas, , & etc., 2019).

2. Experimental research of service system

With the local and global development of computer networks, there is an increase of mixed (several of diverse forms) of heterogeneous network, which connect many sub-networks of capacity, supporting different standards, protocols as well as velocities of network. *HetNet* definition is used in modern heterogeneous wireless computer networks using various types of network nodes for description. Compatibility problems of heterogeneous networks are solved by offering specialized data maintenance solutions (Chang ant etc., 2015, Yang, Chawla and etc., 2012, Qi and etc., 2012), evaluating their combining cases (Shi, 2017), analyzing errors, links (Zhang and etc., 2013, Sajadmanesh and etc., 2016, Yang, Kung, and etc., 2013).

Our solution will help to choose optimal network node detention buffer, by combining networks of different capacity. Absolute majority of companies in local network of *Ethernet* (IEEE 802.3) also realize the wireless sub-network ensuring the accessibility of service for mobile users. Using 802.11n (802.11n-2009 – IEEE Standard for Information technology – Local and Metropolitan area network) standard WLAN (Wireless Local Area Network, IEE 802.11) network node, usually its technical possibilities guarantee 300 Mbps one-channel speed. In the network of company‘s *Gigabit Ethernet* it is ensured with up to 1000 Mbps channel speed. Upon the arrival of queries through these channels into company‘s network server, the efficiency ratio of served channels *r=3.3*. For the optimal work of this node it is sufficient to have buffer of 2-7 applications. If *802.11g WLAN* node is used, then *r=18.5*, and buffer 14-16 applications.

Pattern of computer network node with two different performance channels was used for service system research. Modeled system uses the disciplinary of queue service *FIFO taildrop* (Nzouonta, Ott, Borcea, 2009). It is considered that in the primary state both service system channels and queue are free (t=0), i.e. system is prepared to service the received application immediately. By imitating network node work, length of sequence of moments between the appearance of packaged in the node were generated $\tau_0, \tau_1, \ldots \tau_n$ and length sequence of the package service $x_0, x_1, \ldots x_n$. Using these sequences, the characteristics of package service are being calculated in accordance with the distributions and services procedures of the sequences‘elements. Modeled network flow is generalized stochastically bounded burstiness *gSBB* (hereinafter *gSBB*). This flow for all $t \geq 0$ as well as all $x \geq 0$ satisfy the inequality $P\{\hat{A}(t,\rho) > x\} \leq f(x)$, when ρ is the upper limit of the flow, *f* – coverage function which is not increasing and $f(x) \geq 0$ for all $x \geq 0$, $A(0,t)$ – incoming flow, satisfying the inequality $A(0,t) \ll \langle f, \rho \rangle$ (Jiang, Yin, Liu, Jiang 2009). It is shown that if *gSBB* network flow demonstrates stationary and ergodicity peculiarities, then $P\{\hat{A}(t,\rho) > x\}$ in any moment of times has the upper limit in constant section of queue length, i.e. $P\{\hat{A}(\infty,\rho) > x\}$ in virtual one channel system with constant ρ. Here $\hat{A}(\infty,\rho)$ indicates $\hat{A}(t,\rho)$, when $t \to \infty$ (Jiang, Yin, Liu, Jiang 2009). Length of generated network flow packages is variable and satisfying the requirements of *Ethernet* standard.

(Jiang, Yin, Liu, Jiang 2009) proofed that according to *gSBB* singularity process is modeled according the formula:

$f^{self-similar}(x) = C_\alpha\left(\frac{\rho-m}{\delta}\right)$, when the satisfied inequality is $P\{\hat{A}(t,\rho) > x\} \leq f^{self-similar}(x)$.

Here C_α and *m* are calculated as parameters of $S_\alpha(\beta,\sigma,\mu)$ and *x*. According to *gSBB* pattern, Puason‘s flow is modeled using the formula:

$f^{Poisson}(x) = 1 - (1-\eta) \cdot \sum_{i=0}^{k} \left[\frac{[\eta(i-k)]^i}{i!} e^{-\eta(i-k)}\right]$, when the satisfied inequality is

$P\{\hat{A}(t,\rho) > x\} \leq f^{Poison}(x)$. Here $\eta = \frac{\lambda \underline{S}}{\rho}$ and $k = \frac{x}{\underline{S}}$, where $\underline{S}$ is the average package length.

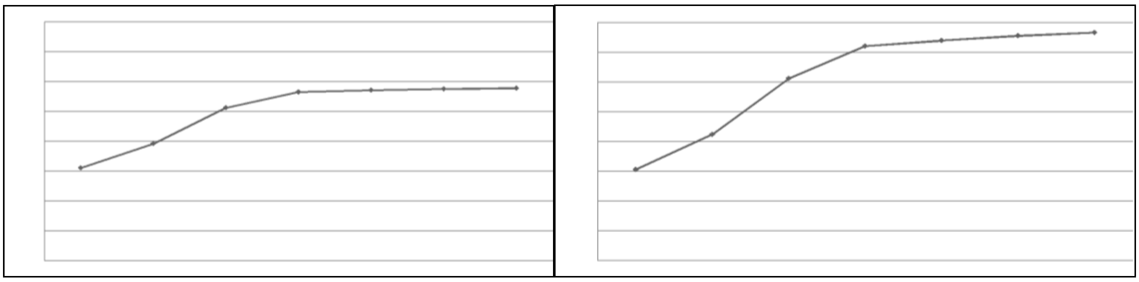

Fig. Average number of queries in the system changing the length of buffer, when *r=5* (left picture) and *r=15 (right picture)*.

The received findings of simulation of system with one efficient (fast) and one inefficient channel (slow) confirmed that 5-10 application buffer is enough for 802.11n and *Gigabit Ethernet* channels (*r=3.3*). When 802.11 n and *Gigabit Ethernet* channels (*r=18.5*) are used then 10-15 application buffer is enough for optimal system work.

3. Conclusions

1. When the flow intensity ρ is close to zero, optimal buffer size varies linearly relative to *r* $K_{opt} = 1{,}12 \cdot r - 2$.

2. When the flow intensity ρ is close to 1, optimal buffer size is calculated using formula $K_{opt} = 1{,}12 \cdot r^2 + o(r^2)$.

3. In case of high capacity of channels to ratio r, numbers of application becomes the same as using one-channel system, therefore in which case it is recommended to refuse the second inefficient channel.

References

1. Bilel, B. R., Navid, N., & Bouksiaa, M. S. M. (2012, October). Hybrid cpu-gpu distributed framework for large scale mobile networks simulation. In Proceedings of the 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications (pp. 44-53). IEEE Computer Society.

2. Chang, S., Han, W., Tang, J., Qi, G. J., Aggarwal, C. C., & Huang, T. S. (2015, August). Heterogeneous network embedding via deep architectures. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 119-128). ACM.

3. Efrosinin, D. V., & Rykov, V. V. E. (2008). On performance characteristics for queueing systems with heterogeneous servers. Automation and Remote Control, 69(1), 61-75.

4. Efrosinin, D., & Sztrik, J. (2011). Performance analysis of a two-server heterogeneous retrial queue with threshold policy. Quality Technology & Quantitative Management, 8(3), 211-236.

5. Erlong, A. K. (1909). The theory of probabilities and telephone conversations. Nyt Tidsskrift Mat. Vol. 20, pp. 33-39.

6. Gelenbe, E., Pujjole G. (1999). Introduction to queuing networks, second edition. John Willey & Sons, pp. 258

7. Hetherington, T. H., Rogers, T. G., Hsu, L., O'Connor, M., & Aamodt, T. M. (2012, April). Characterizing and evaluating a key-value store application on heterogeneous CPU-GPU systems. In Performance Analysis of Systems and Software (ISPASS), 2012 IEEE International Symposium on (pp. 88-98). IEEE.

8. Iversen, V. B. (2011). Teletrafic Engineering And Network Planning. Technical University of Denmark, pp. 567.

9. Kadjo, D., Ayoub, R., Kishinevsky, M., & Gratz, P. V. (2015, June). A control-theoretic approach for energy efficient CPU-GPU subsystem in mobile platforms. In Proceedings of the 52nd Annual Design Automation Conference (p. 62). ACM.

10. Kaklauskas, L., Sakalauskas, L., & Denisovas, V. (2019). Stalling for solving slow server problem. RAIRO-Operations Research, 53(4), 1097-1107.

11. Mokaddis, G. S., Matta, C. H., & El Genaidy, M. M. (1998). On Poisson queue with three heterogeneous servers. International journal of information and management sciences, 9(4), 53-60.

12. Qi, G. J., Aggarwal, C., & Huang, T. (2012, April). Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces. In Proceedings of the 2012 SIAM International Conference on Data Mining (pp. 528-539). Society for Industrial and Applied Mathematics.

13. Rubinovitch, M. (1985). The slow server problem: a queue with stalling. Journal of Applied Probability, 22(04), 879-892.

14. Rykov, V. V. E., & Efrosinin, D. V. (2009). On the slow server problem. Automation and Remote Control, 70(12), 2013-2023.

15. Rykov, V., & Efrosinin, D. (2004). Optimal control of queueing systems with heterogeneous servers. Queueing Systems, 46(3), 389-407.

16. Sajadmanesh, S., Rabiee, H. R., & Khodadadi, A. (2016, August). Predicting anchor links between heterogeneous social networks. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 158-163). IEEE.

17. Sakalauskas, L. (2000). Masinio aptarnavimo teorija. VGTU, „Technika“, pp. 156.

18. Sankaranarayanan, K., Larsen, E. R., van Ackere, A., & Delgado, C. A. (2010, December). Genetic algorithm based optimization of an agent based queuing system. In Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on (pp. 1344-1348). IEEE.

19. Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2017). A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 29(1), 17-37.

20. Xiaolong, J., Geyong M. (2009). Modelling and Analysis of Priority Queuing Systems with Multi-Class Self-Similar Network Traffic: A Novel and Efficient Queue-Decomposition Approach. IEEE Transactions on Communications; vol. 57, No. 5, pp. 1444-1452

21. Yang, C. L., Kung, P. H., Li, C. T., Chen, C. A., & Lin, S. D. (2013, December). Sampling Heterogeneous Networks. In Data Mining (ICDM), 2013 IEEE 13th International Conference on (pp. 1247-1252). IEEE.

22. Yang, Y., Chawla, N., Sun, Y., & Hani, J. (2012, December). Predicting links in multi-relational and heterogeneous networks. In Data Mining (ICDM), 2012 IEEE 12th International Conference on (pp. 755-764). IEEE.

23. Yue, D., Yue, W., Yu, J., & Tian, R. (2009). A heterogeneous two-server queuing system with balking and server breakdowns. In Eight International Symposium on Operations Research and its Applications (ISORA’09). Zhangjiajie, Chaina.

24. Zhang, W., Wang, S., Yang, Y., & Wang, Q. (2013, November). Heterogeneous network analysis of developer contribution in bug repositories. In Cloud and Service Computing (CSC), 2013 International Conference on (pp. 98-105). IEEE.

Received: 2 November 2021
Accepted: 27 January 2022