Performance Analysis of the M/M/2/2 System with Heterogeneous Servers

Hakyong KIM*

R&D Center, Samsung Networks

The approach of using multiple heterogeneous servers instead of using a single server or multiple homogeneous servers has been investigated in different point of views as varied computing or switching capabilities in a system come to be requested [1-2]. The research interest particularly has been focused on a two-server system due to its analytical simplicity as well as the efficiency in the performance improvement.

In the system composed of two heterogeneous servers, the system performance is dependent of the service ratio between two servers. In this paper, we define this service ratio between two servers as the service ratio, β, and investigate the system in terms of β. In this study, we confine our research interest into the queue-less heterogeneous two-server system. Namely, an M/M/2H/2 system.

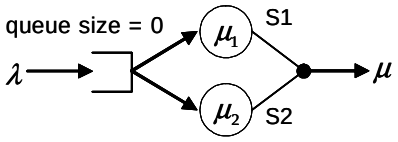

Fig. 1. Queueing Model of the M/M/2H/2 system.

The queueing model of the M/M/2H/2 system is shown in Fig. 1. As shown in the figure, the system employs two heterogeneous servers S1 and S2 and uses zero queue space. Here, we assume that the service rate of server S1 is faster than that of server S2. That is, $\mu_1 > \mu_2$.

When a new packet arrives at this system, the packet is assigned to fast server S1 with preference. If server S1 is busy, then the newly-arrived packet is assigned to slow server S2. If both servers are busy, then the packet is blocked and lost. From this service policy, we have the state transition diagram for the M/M/2H/2 system as shown in Fig. 2.

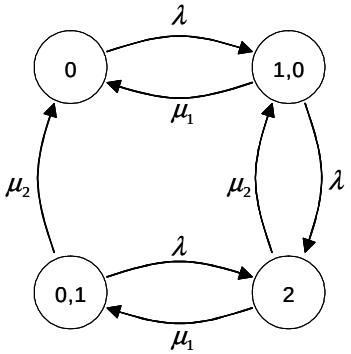

Fig. 2. State transition diagram of the M/M/2H/2 system.

The state of the M/M/2H/2 system in Fig. 2 is defined to be the system size, N, when a new packet arrives. When $N=1$, we need to indicate which server is busy, server S1 or server S2. In Fig. 2, the doublet (N_1, N_2) represents the number of packets in server S1 and server S2, while the single numeral in the rest states indicates the total number of packets in the system. Here, let P_k be the probability of state k ($k \in \{0, (0,1), (1,0), 2\}$). Then, we can derive 4 equations from the state transition diagram of Fig. 2. in terms of 4 state probabilities (P_0, P_{01}, P_{10}, and P_2), packet arrival rate λ, system service rate $\mu = \mu_1 + \mu_2$, utilization factor $\rho = \lambda/(\mu_1 + \mu_2)$, and service ratio $\beta = \mu_2 / \mu_1$. Solving these 4 equations together with the identity equation of $P_0 + P_{01} + P_{10} + P_2 = 1$, we have

$$P_2 = \frac{\rho^2\left[(1+\rho)\beta^2 + (1+2\rho)\beta + \rho\right]}{\rho(1+\rho)^2\beta^2 + (1+2\rho)(1+\rho+\rho^2)\beta + \rho^2(1+\rho)} \quad (1)$$

By comparing Eq. (1) with the Erlang B probability P_b of the M/M/2/2 system which employs homogeneous servers, we derive an important result that $P_2 < P_b$ for $\beta_2 < \beta < \beta_1$ where $\beta_1 = 1$ and $\beta_2 = \rho/(1+\rho)$. Furthermore, differentiating Eq. (1) in terms of β, we can show that there exists such β that minimizes P_2. Let P_2^* be the minimum of P_2.

P_2 has two boundary values as β approaches 0 and 1. Let these boundary values P_{2u} and P_{2l}, respectively. Then, for the target Erlang B probability a, we have the following result:

(a) If $P_2^* < a \le P_{2l}$, the desired interval is $\beta_2 < \beta < \beta_1$.
(b) If $P_{2l} < a \le P_{2u}$, the desired interval is $\beta_2 < \beta < 1$.
(c) If $P_{2u} < a$, the desired interval is $0 < \beta < 1$.

According to [3], our findings will be hold for general arrival distributions.

References

[1] N. Gogate and S. Panwar, IEEE Comm. Letters, vol. 3, no. 4, p. 119-121, (1999).
[2] W. Lin and P. Kumar, IEEE Tr.. on Automatic Control, vol. 29, no. 8, p. 696-703, (1984).
[3] S. M. Ross, Stochastic Processes, John Wiley & Sons, (1983).