On the NT-Policy for Discrete-Time Queues

B. Feyaerts, S. De Vuyst, S. Wittevrongel, and H. Bruneel

SMACS Research Group, TELIN Department, Ghent University,
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
{bfeyaert,sdv,sw,hb}@telin.ugent.be

Abstract

We study a discrete-time single-server queueing system operating under the NT-policy. This policy aims at clustering the service of customers in order to reduce the number of server activations and deactivations. A customer arriving in an empty system, will have to wait until one of two thresholds is reached before its service is started. The first threshold — a space threshold — is reached when the queue length reaches N. The second threshold — a time threshold — is reached when the first customer has been waiting for T slots.

Keywords: Queueing theory, NT-policy

In typical queuing systems, low to moderate load conditions will cause frequent activation and deactivation of the service unit. This frequent transition may pose a severe overhead, e.g. with machines that undergo a costly initialisation procedure before the service unit is fully operational after a period of idleness. This cost can be of various natures: time, power consumption, peak current, ... In such cases, it can be beneficial to purposely delay the service of customers in order to create service clusters.

This can be achieved by applying a threshold policy, i.e. a mechanism that stalls the service unit activation until some threshold has been reached. The N-policy is one of the most intuitive threshold policies and was first presented in [1]. Under this policy, once the system becomes empty and the service unit is deactivated, it will only be reactivated when N customers/jobs have accumulated again. The main drawback of the N-policy is that very low load conditions can cause unacceptable delays.

The NT-policy was constructed to counter this drawback by imposing a time threshold T. The service unit will be reactivated when either the queue has reached a length of N customers, or if the first customer has been waiting in the queue for a time T, whichever happens first.

Threshold policies generally introduce a three-phase cyclic pattern. When a first customer arrives in an empty system, the system proceeds to an accumulating phase until the threshold policy activates the service unit. The system will then start serving the

customers exhaustively until it becomes empty again. Thus, we identify three subsequent phases, i.e. *empty*, *accumulating* customers and *serving* customers. The term cycle is used to refer to a series of the three consecutive phases, starting with the empty phase.

In this research, we analyze a discrete-time single-server queueing system operating under the NT-policy. We assume a Bernoulli arrival process with arrival rate λ and deterministic service times of 1 slot. Our analysis focuses on the server phase durations and the customer delay distribution.

To model the system, we introduce three random variables ϕ_k, t_k and u_k, for a random slot k. The variable ϕ_k denotes the phase in which the system resides during slot k. The random variable t_k corresponds to the sojourn time of the first customer in the queue at the end of slot k and is only meaningful during the accumulating phase. Finally, u_k shows the system content at the beginning of slot k. We can combine these three variables in a system state vector that describes the system state at a random slot k.

Note that the system state space is finite, since t_k and u_k can only increase during the accumulating phase until either one of the thresholds is met. This allows us to fully explore the state space and calculate the probability of each possible state. Taking into account that the system state during the first slot of the accumulating phase is identical every cycle and appears exactly once per cycle, we can transform the system state probabilities to expressions that refer to the fraction of a cycle that the system is in that state. This transformation simplifies our analysis and allows us to write expressions more concisely than with the mere probabilities.

In order to find the phase and cycle duration distribution, it can be helpful to depict the probabilities, or the corresponding cycle fractions, in a probability graph. Phase durations then correspond to a walk through the complete subgraph of states of the same phase and the cycle duration corresponds to a walk through the whole graph, where each phase is visited.

For the customer delay, we pick and tag a random customer that arrives during a certain phase and calculate the number of slots it takes until that customer leaves the system. Repeating this for all three possible phases, we get the complete customer delay distribution.

For large values of N and T, the calculation of the probabilities or the corresponding cycle fractions can become quite time consuming. In such cases, where also λT and N differ significantly, the means of the phase durations, the system content and the customer delays can be approximated elegantly. The system content distribution will show triangular characteristics that can be exploited.

Numerical tests have shown that in terms of customer delay, the NT-policy clearly outperforms the N-policy for $\lambda \leq N/T$. For larger values of λ, both policies have about the same performance.

References

[1] M. Yadin and P. Naor, Queueing systems with a removable service station, *Operational Research Quarterly*, vol. 14, no. 4, pp. 393–405 (1963)