Mean Value Formulae for the Queues M/G/1-SPT & M/G/1-SRPT: A New Method of Derivation Using Differential Equations

H Y Ranjit Perera[1]

1. Introduction

Widely used scheduling strategies such as FIFO (First In First Out), Random and LIFO (Last In First Out) lead to mean waiting times that do not depend on the service time of the jobs. But other strategies such as SPT (Shortest Processing Time first) and SRPT (Shortest Remaining Processing Time first) lead to service time dependant mean waiting times and such strategies are thus called biased strategies [1]. Among all known strategies, SRPT produces the smallest overall mean waiting time and is thus also called the optimal strategy [2]. Among all the non-preemptive strategies SPT turns out to be the optimal one. Derivation of the implicit Laplace transforms for related functions in M/G/1-SPT and SRPT has been carried out [3] and also the mean values are obtained through building the derivatives of the Laplace transform. In [6] a different approach has been used to obtain the mean value formula considering SPT and SRPT to be limiting cases of non-preemptive and preemptive-resume HOL (Head Of Line) strategies respectively. This paper introduces a new approach using differential equations leading to all the relevant mean values.

2. The Queue M/G/1-FIFO

The queue M/G/1-FIFO has been extensively investigated and the related Pollaczek-Khinchin (P-K) mean value formula is being widely used. Usually a complicated imbedded Markov-Chain approach is used to derive this equation. An alternative derivation using Little's Law [4] is repeated below. It should be noted, that this approach leads only to the mean values and thus is not superior to the imbedded Markov-Chain approach. Nevertheless, this method turns out to be of some educational importance and can also be extended to some other problems as done later in this paper.

The arrival and departure rates of the queue are denoted by λ and μ respectively. The system load is denoted by ρ and is equal to (λ/μ). We consider a stationary queue under the condition $\rho < 1$. We also make use of the fact that a new arrival finds the queue busy with probability ρ. In order to analyse the properties of the queue we trace the fate of an arbitrarily selected job that we name here a test job. Let us consider the arrival of a test job, that meets N_q (= 0, 1, 2, 3, ...) jobs waiting in the queue. Each of these jobs have an arbitrary service time that is independent of each other but has the same distribution with mean equal to $1/\mu$. The mean value of N_q, denoted by $E[N_q]$, is related to the mean waiting time W through Little's law:

$$E[N_q] = \lambda W \tag{2.1}$$

In case the test job meets the busy system, there is exactly one job in service, which is not counted in N_q. In other words, the test arrival encounters a partially served job with a probability of ρ. If we denote the random value of the service time by x with mean $1/\mu$ and coefficient of variation C_b and its residual time by r the expected value $E[r]$ is given by [5]

$$E[r] = E[x^2]/(2E[x]), \quad \textit{where} \quad E[x] = 1/\mu \quad \textit{and} \quad E[x^2] = (1 + C_b^2)/\mu^2 \tag{2.2}$$

The waiting time of the test job is obtained by summing up service times of the jobs in waiting and the remaining service time of the job in service. As all the service times are independent of each other and also

[1] Prof. Ranjit Perera is a Professor in Electrical Engineering at University of Moratuwa (hyrp@elect.mrt.ac.lk) and is currently working as the Dean, Faculty of Engineering at the General Sir John Kotelawala Defence University, Ratmalana (hyrp@kdu.ac.lk).

All three of these contributions add together to make δW and this relationship is given below:

$$\delta W = 2\lambda x W \quad p(x)\delta x + \delta W \quad \lambda \int_0^x tp(t)dt \tag{3.1}$$

By letting $\delta x \to 0$ the following differential equation is obtained:

$$\left[1 - \lambda \int_0^x tp(t)dt\right] \frac{dW}{dx} = 2\lambda x \quad p(x)W \tag{3.2}$$

This can be solved for W and if we denote W for the limiting case of $x \to 0$ through W_0 the solution can be presented in the form

$$W = \frac{W_0}{\left[1 - \lambda \int_0^x tp(t)dt\right]^2} \tag{3.3}$$

In order to find W_0 two situations given below are considered:

1. The test job meets the system in idle state. This occurs with a probability of $(1-\rho)$ and the waiting time corresponding to this situation is zero.
2. The test job meets the busy system. This occurs with a probability of ρ and the waiting time of the test job with $x \to 0$ is equal to the mean residual time $E[r]$ of the job being served.

These two situations can be combined to get W_0 leading to

$$W_0 = \rho \quad E(r) = \rho \frac{1 + C_b^{\ 2}}{2\mu} \tag{3.4}$$

Equations (3.3) and (3.4) together give the complete solution for the mean waiting time of x-job. The overall mean waiting time, here denoted by W_{all}, can be obtained by performing an integration for all possible values of x as shown below:

$$W_{all} = \int_0^\infty W p(x)dx \tag{3.5}$$

4. The Queue M/G/1-SRPT

The SRPT strategy is a pre-emptive resume strategy and jobs in service can be interrupted in favour of the new-coming shorter jobs. Interrupted jobs queue up, for an intermediate waiting time till they are later allowed to resume service. The intermediate waiting times can also be called interruption times as they are the waiting times caused by interruptions. Thus, relatively long jobs can even be interrupted several times and consequently experience an equal number of intermediate waiting times. Here, it is more meaningful to derive the mean system time of an x-job, instead of mean waiting time, comprising all the waiting times and the service time. However, the derivation is made simple if the system time is broken into the two components *first waiting time* with mean W and *residence time* with mean R. The first waiting time is the time spent by the

test-job since its arrival until it is first taken for service. At the end of the first waiting time the residence time starts and terminates once the job departs after service has been completed. Thus, the residence time includes all the intermediate waiting times if any and the complete service time.

Let us first study the residence time. We consider a test-job whose rest service time is r $(r \le x)$ and denote the mean time taken to reduce the remaining service time from r to $r-\delta r$ by $\Delta R(r)$. It is observed that $\Delta R(r) > \delta r$ due to possible interruptions. It is highlighted that ΔR is a function of r and $\Delta R(r+\delta r) > \Delta R(r)$. We also denote $\Delta R(r+\delta r) - \Delta R(r)$ by $\delta(\Delta R)$. This difference is due to the interruptions caused by

1. New arrivals whose service time falls in $[r, r+\delta r)$ and occur during $\Delta R(r+\delta r)$. They interrupt jobs with remaining time $r+\delta r$ but not the jobs with remaining time r and thus contribute to $\delta(\Delta R)$.
2. New arrivals whose service time falls in $[0, r+\delta r)$ and occur during $\delta(\Delta R)$.

As done in section 3 the contributions due to above interruptions can also be quantified. The rate at which work belonging to first and second categories of interruptions arrive simplify to, after dropping the insignificant terms, $\lambda r p(r)\delta r$ and $\lambda \int_0^r tp(t)dt$ respectively. The corresponding mean amount of work is obtained by multiplying these rates by the respective mean durations and they work out to $\lambda r p(r)\delta r \Delta R(r)$ and $\lambda \int_0^r tp(t)dt \cdot \delta(\Delta R)$ respectively. They can be now added together to form $\delta(\Delta R)$ leading to the following equation:

$$\delta(\Delta R) = \lambda \quad r\ p(r)\delta r \qquad \Delta R(r) + \lambda \int_0 tp(t)d\delta(\Delta R) \tag{4.1}$$

This again leads to the differential equation

$$[1 - \lambda \int_0 tp(t)dt]\frac{d(\Delta R)}{dr} = \lambda \quad r\ p(r)\Delta R \tag{4.2}$$

Using the fact that ΔR tends to δr for $r \to 0$ this differential equation is solved for ΔR to give

$$\Delta R = \frac{\delta r}{1 - \lambda \int_0^r tp(t)dt} \tag{4.3}$$

The mean residence time R of a x-job is obtained by integrating ΔR from $r = 0$ to $r = x$:

$$R = \int_0^x \frac{dr}{1 - \lambda \int_0^r tp(t)dt} \tag{4.4}$$

Let us now get on to the mean of the first waiting time of an x-job denoted by W. As already discussed in section 3 a job of service time $(x + \delta x)$ has a mean first waiting time slightly larger than W and

is denoted by $(W+\delta W)$. The contributions that are responsible for δW are caused by four distinct categories of jobs:

1. Waiting $(x, x+\delta x]$ -jobs at the arrival of the test-job.
2. New $[x, x+\delta x)$ -jobs that arrive during the mean first waiting time W
3. New $[0, x+\delta x)$ -jobs that arrive during the mean duration δW
4. Residing jobs with a remaining service time in the interval $(x, x+\delta x]$ at the arrival of the test-job.

The mean amount of work due to jobs in categories 1, 2 and 3 work out to, after dropping the insignificant terms, $\lambda x p(x)\delta x W$, $\lambda x p(x)\delta x W$ and $\lambda\int_0^x tp(t)\,dt \quad \delta W$ respectively. All the jobs whose original service time has been larger than x become, at a later stage, jobs with remaining service time in the interval $(x, x+\delta x]$ and thus contribute to category 4. The arrival rate of such jobs is given by $\lambda[1-\int_0^x p(t)dt]$ and the corresponding residence time is given by $\Delta R(x)$. Using Little's law we get the mean number of jobs in category 4 to be $\lambda[1-\int_0^x p(t)dt]\cdot\Delta R(x)$. Each one of these jobs have a remaining service time equal to x and this work has priority over the new-coming $(x+\delta x)$-job. The mean amount of work corresponding to this is $\lambda[1-\int_0^x p(t)dt]\cdot x\Delta R(x)$. Now we can add all these four contributions together to form δW

$$\delta W = 2\lambda x p(x)\delta x W + \lambda\int_0^x tp(t)d\delta W + \lambda[1-\int_0^x p(t)dt]x\Delta R(x) \tag{4.5}$$

Equation (4.3) gives $\Delta R(r)$ and this result can be used to replace $\Delta R(x)$ in the above equation.

$$[1-\lambda\int_0^x tp(t)dt]^2\delta W = [1-\lambda\int_0^x tp(t)dt]2\lambda x\mathrm{p(x)W}\delta\mathrm{x} + \lambda x[1-\int_0^x p(t)dt]\delta\mathrm{x} \tag{4.7}$$

The following differential equation results from this:

$$[1-\lambda\int_0^x tp(t)dt]^2\frac{dW}{dx} = [1-\lambda\int_0^x tp(t)dt]2\lambda x\mathrm{p(x)W} + \lambda x[1-\int_0^x p(t)dt] \tag{4.8}$$

The first term on the right hand side can be brought to the left side and combined together using the rule for differentiation of a product giving the more simplified relationship:

$$d\{[1-\lambda\int_0^x tp(t)dt]^2 W\} = \lambda x[1-\int_0^x p(t)dt]dx \tag{4.9}$$

Unlike in the case of SPT x -jobs with $x\to 0$ have zero first waiting time as no other job has priority over them under SRPT. This means $W_0 = 0$ for this case. Using this fact und doing an integration by parts on the right hand side of the equation (4.9) we get the following result for W :

$$W = \frac{\lambda}{2} \cdot \frac{\int_0^x t^2 p(t)dt + x^2[1 - \int_0^x p(t)dt]}{[1 - \lambda \int_0^x tp(t)dt]^2} \tag{4.10}$$

With this the complete mean system time of an x-job in M/G/1-SRPT is obtained and given by $W + R$ with W and R as in equations (4.10) and (4.4) respectively. Similar to equation (3.5) the overall mean system time S_{all} works out to:

$$S_{all} = \int_0^\infty (W + R)\, p(x)dx \tag{4.11}$$

8. Concluding Remarks

Mean waiting times in queues using biased strategies such as SPT and SRPT depend on the service time and are continuous functions of the service time. The original work on M/G/1-SPT and SRPT done by Schrage and Miller [1] to obtain the implicit Laplace transforms of the related functions are also used to get mean values but involve fairly long calculations. A new approach using differential equations to obtain the same mean values are introduced in this paper. This method does not involve Laplace transforms and thus gives only the relevant mean values. The method of derivation is comparatively simple and is also of some educational importance and may find further applications in other problems.

Acknowledgement

The author gratefully acknowledges the contribution made by Prof. Friedrich Schreiber to the topic of optimal strategies SPT and SRPT. This work has also evolved as a result of his initiation. Contribution made by Prof. Carmelita Görg, University of Bremen, by giving critical comments is also especially acknowledged.

References

[1] Ruschitzka, M., Farby, R. S., "A unifying approach to scheduling". Com. ACM 20, 7, 1977, pp 469-477.

[2] Schrage, L. E., "A proof of the optimality of the shortest remaining processing time discipline". Opns. Res. 16, 1968, pp 687-690.

[3] Schrage, L. E., Miller, L. W., "The queue M/G/1 with the shortest processing time discipline". Opns. Res. 14, 1966, pp 670-684.

[4] Little, J. D. C., "A proof for the queuing formula: $L = \lambda W$ ", Oper. Res. 9, 1961, pp 383-387.

[5] Kleinrock, L., "Queueing systems, volume I: Theory". Wiley and Sons, New York, London 1975.

[6] Conway, R. W., W-L. Maxwell, L.W. Miller, "Theory of scheduling", Addision-Wesley, Reading/Mass., 1967.

[7] Perera, R., "Beiträge zur Theorie von Wartesystemen mit den Optimalstrategien SPT und SRPT", Ph. D. Thesis, RWTH Aachen 1989.

[8] Görg, C., "WARTERAUM M/: Die SRPT-Optimalstrategie im Vergleich mit der Zeitscheibenstrategie unter Berücksichtigung von Verwaltungszeiten", Ph. D. Thesis, RWTH Aachen 1983.

[8] Görg, C., Pham, X. H., "Improving mean delay in data communication networks by new combined strategies based on the SRPT-principle", Proc. 11. ITC, Kyoto/Japan, Sept. 1985.