Appendices

Spectrum Requirement Planning in Wireless Communications: Model and Methodology for IMT-Advanced Edited by H. Takagi and B. H. Walke © 2008 John Wiley & Sons, Ltd. ISBN: 978-0-470-98647-9

Appendix A

Derivation of Formulas by Queueing Theory

Hideaki Takagi

In this appendix, we derive the basic formulas used in the methodology for determining the capacity requirement as shown in Table A.1. These formulas are derived by the theory of queues. Erlang-B formula for the blocking probability in a loss system, Erlang-C formula for the wait probability in a delay system, and Cobham's formula for the average waiting time in an M/G/1 nonpreemptive priority queue appear in introductory textbooks of queueing theory such as Cooper (1991), Gross and Harris (1998), Kleinrock (1975), and Kleinrock (1976) as well as in monographs on teletraffic engineering such as Fujiki and Gambe (1980) and Syski (1986). We refer to Wolff (1989) for the proofs of fundamental properties of queues. The multidimensional Erlang-B formula for the blocking probability in a loss system with multiple classes of calls and multiple server occupation is an example of so-called product-form solution to a network of queues that can be modeled by a reversible Markov process (Iversen). However, we present its analysis and computational algorithm in an elementary manner by following Kaufman (1981).

A basic queueing model consists of an arrival stream of customers, a waiting room and a set of servers. Various queueing models are conveniently described by the so-called *Kendall's notation* 'A/B/s/K' as explained in Gross and Harris (1998, p. 8) and Kleinrock (1975, p. 399). In this notation, the first parameter 'A' indicates the arrival process such as 'M' for the Poisson process. The second parameter 'B' indicates the service time distribution such as 'M' for the exponential distribution, 'D' for the deterministic, i.e. constant service time, and 'G' for the general distribution. The third parameter s denotes the number of servers. The fourth parameter K denotes the maximum number of customers that can be accommodated in the whole system, i.e. both in the waiting room and servers. The fourth parameter may be omitted if there is no restriction on the capacity of the waiting room; this model is a delay

Spectrum Requirement Planning in Wireless Communications: Model and Methodology for IMT-Advanced Edited by H. Takagi and B. H. Walke ISBN: 978-0-470-98647-9

system. If $K = s$, the model is a loss system, because there is no waiting room. The first three parameters seem to be quite standard. However, the reader should note that the fourth parameter is used with different meaning in some books, e.g. Walke (2002, p. 1043).

Table A.1 Sections where formulas of queueing theory are used.

Formula	System	Traffic	Section
Erlang-B formula for a loss system	IMT-2000	circuit-switched	3.3.2
Erlang-C formula for a delay system	IMT-2000	packet-switched	3.3.2
Multidimensional Erlang-B formula	IMT-Advanced	circuit-switched	4.3.4
M/G/1 nonpreemptive priority queue	IMT-Advanced	packet-switched	4.3.5

A.1 Erlang-B Formula for a Loss System

Let us consider an M/M/s/s loss system depicted in Figure A.1. There are s servers and no waiting room. Calls arrive in a Poisson process with rate λ. The service time of each call has exponential distribution with mean $1/\mu$. Calls that arrive when all servers are busy are blocked and lost.

We define the state of the system by the number of calls present in the system. The state space is finite. The state follows a birth-and-death process, for which the state transition rate diagram is shown in Figure A.2.

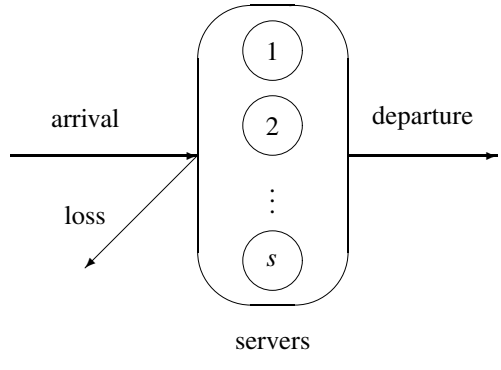


Figure A.1 A loss system.

Let p_k be the probability that there are k calls in the system at an arbitrary time in equilibrium, where $k = 0, 1, 2, \ldots, s$. Then the set of balance equations for $\{p_k; 0 \le k \le s\}$ is given by

$$\begin{aligned} \mu p_1 &= \lambda p_0 \\ \lambda p_{k-1} + (k+1)\mu p_{k+1} &= (\lambda + k\mu) p_k, \quad 1 \le k \le s-1 \\ \lambda p_{s-1} &= s\mu p_s. \end{aligned} \tag{A.1}$$

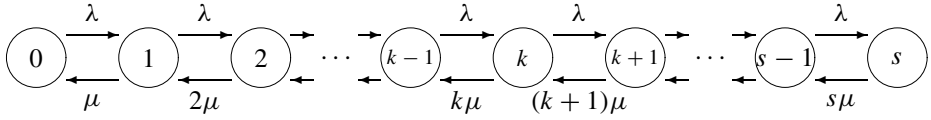


Figure A.2 State transition rate diagram for an M/M/*s*/*s* loss system.

This is equivalent to the set

$$k\mu p_k = \lambda p_{k-1}, \quad 1 \le k \le s \tag{A.2}$$

which gives

$$p_k = \frac{\lambda}{k\mu} p_{k-1} = \frac{\lambda^2}{k(k-1)\mu^2} p_{k-2} = \cdots = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k p_0, \quad 1 \le k \le s. \tag{A.3}$$

Writing $a := \lambda/\mu$ for the offered traffic, and finding p_0 from the normalization condition

$$\sum_{k=0}^{s} p_k = 1, \tag{A.4}$$

we obtain the probability distribution for the number of calls in the system

$$p_k = p_0 \frac{a^k}{k!} = \frac{a^k/k!}{\sum_{j=0}^{s}(a^j/j!)}, \quad 0 \le k \le s. \tag{A.5}$$

This is the *truncated Poisson distribution*.

Calls that arrive when all servers are used are blocked. From the *PASTA* (Poisson arrivals see time averages) property (Wolff 1989, pp. 293–297), the probability that there are k calls in the system immediately before an arrival time equals the probability that there are k calls in the system at an arbitrary time given in Equation (A.5). Therefore, the *blocking probability* is given by

$$p_s = \frac{a^s/s!}{\sum_{j=0}^{s}(a^j/j!)} := E_{\mathrm{B}}(s, a). \tag{A.6}$$

This is called the *Erlang-B formula* or *Erlang's loss formula*. This formula was first derived by A. K. Erlang in 1917.[1]. This result is used in the methodology of required capacity calculation for the circuit-switched services in IMT-2000 systems in Section 3.3.2.

We note the relation

$$E_{\mathrm{B}}(0, a) = 1; \quad E_{\mathrm{B}}(s, a) = \frac{a E_{\mathrm{B}}(s-1, a)}{s + a E_{\mathrm{B}}(s-1, a)}, \quad s \ge 1 \tag{A.7}$$

which was very useful to calculate $E_{\mathrm{B}}(s, a)$ for $s = 1, 2, \ldots$, in this order recursively for a given value of a in the era of B. C. (before computer).

[1] See Brockmeyer *et al.* (1948) for the life and works of A. K. Erlang.

Erlang-B formula is derived in Cooper (1991, p. 80), Fujiki and Gambe (1980, p. 47), Gross and Harris (1998, p. 80), Kleinrock (1975, p. 106), and Syski (1986, p. 147).

It is noteworthy that Equations (A.5) and (A.6) are valid even when the service time has general distribution, i.e. for an M/G/s/s loss system. In such a case, $1/\mu$ in the above formulas is replaced by the average service time. This property is called the *insensitivity* (Wolff 1989, pp. 271–273).

A.2 Erlang-C Formula for a Delay System

Let us consider an M/M/s delay system depicted in Figure A.3. There are s servers and a waiting room of infinite capacity. Calls arrive in a Poisson process with rate λ. The service time of each call has exponential distribution with mean $1/\mu$. Calls that arrive when all servers are busy wait in the waiting room.

We define the state of the system by the number of calls present in the system. The state space is infinite. The state follows a birth-and-death process, for which the state transition rate diagram is shown in Figure A.4.

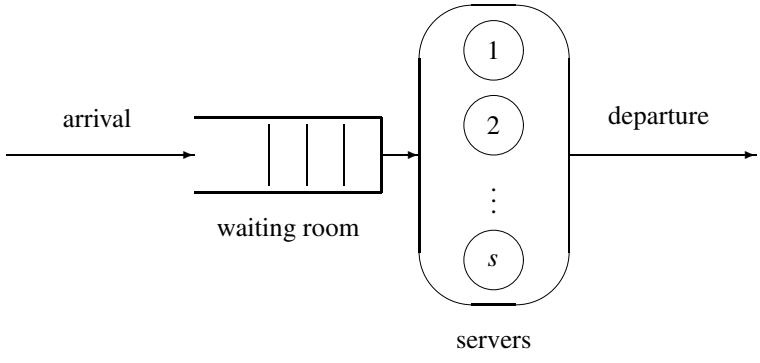


Figure A.3 A delay system.

Let p_k be the probability that there are k calls in the system at an arbitrary time in equilibrium, where $k = 0, 1, 2, \ldots$. Then the set of balance equations for $\{p_k; k \geq 0\}$ is given by

$$\begin{aligned} \mu p_1 &= \lambda p_0 \\ \lambda p_{k-1} + (k+1)\mu p_{k+1} &= (\lambda + k\mu) p_k, \quad 1 \leq k \leq s-1 \\ \lambda p_{k-1} + s\mu p_{k+1} &= (\lambda + s\mu) p_k, \quad k \geq s. \end{aligned} \tag{A.8}$$

This is equivalent to the set

$$k\mu p_k = \lambda p_{k-1}, \quad 1 \leq k \leq s; \quad s\mu p_k = \lambda p_{k-1}, \quad k \geq s \tag{A.9}$$

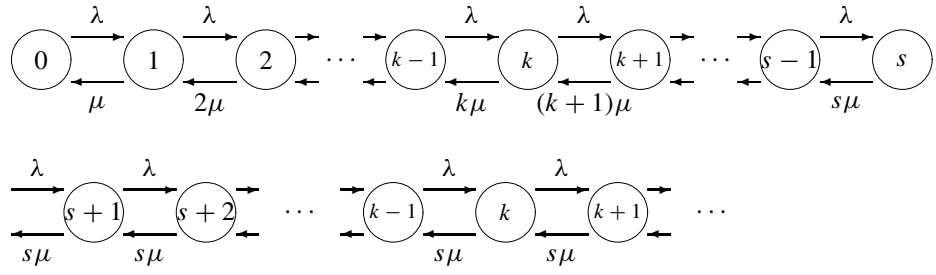


Figure A.4 State transition rate diagram for an M/M/s delay system.

which gives

$$p_k = \frac{\lambda}{k\mu} p_{k-1} = \frac{\lambda^2}{k(k-1)\mu^2} p_{k-2} = \cdots = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k p_0, \quad 1 \le k \le s \tag{A.10}$$

$$\begin{aligned} p_k &= \frac{\lambda}{s\mu} p_{k-1} = \left(\frac{\lambda}{s\mu}\right)^2 p_{k-2} = \cdots = \left(\frac{\lambda}{s\mu}\right)^{k-s} p_s \\ &= \frac{(\lambda/\mu)^k}{s!s^{k-s}} p_0 = \frac{s^s}{s!}\left(\frac{\lambda}{s\mu}\right)^k p_0, \quad k \ge s. \end{aligned} \tag{A.11}$$

Writing $a := \lambda/\mu$ for the offered traffic, we determine p_0 from the normalization condition

$$\sum_{k=0}^{\infty} p_k = 1. \tag{A.12}$$

When $a/s < 1$ (the stability condition), we obtain

$$\frac{1}{p_0} = \sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{s^s}{s!} \sum_{k=s}^{\infty} \left(\frac{a}{s}\right)^k = \sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s}{(s-1)!(s-a)}. \tag{A.13}$$

Thus we obtain the probability distribution for the number of calls in the system

$$p_k = \begin{cases} p_0 \dfrac{a^k}{k!}, & 0 \le k \le s \\ p_0 \dfrac{s^s}{s!}\left(\dfrac{a}{s}\right)^k = p_s \left(\dfrac{a}{s}\right)^{k-s}, & k \ge s. \end{cases} \tag{A.14}$$

Calls that arrive when all servers are used must wait in the waiting room. Owing to the PASTA property, the wait probability is given by

$$\begin{aligned} P\{W > 0\} &= \sum_{k=s}^{\infty} p_k = p_s \sum_{k=s}^{\infty} \left(\frac{a}{s}\right)^{k-s} = \frac{p_s s}{s-a} = \frac{p_0 a^s}{(s-1)!(s-a)} \\ &= \frac{a^s/(s-1)!(s-a)}{\sum_{j=0}^{s-1} a^j/j! + a^s/(s-1)!(s-a)} := E_C(s,a). \end{aligned} \tag{A.15}$$

This is called the *Erlang-C formula* or *Erlang's delay formula*. This formula was also derived by A. K. Erlang in 1917.

We may note the relationship

$$E_{\mathrm{C}}(s,a) = \frac{sE_{\mathrm{B}}(s,a)}{s-a+aE_{\mathrm{B}}(s,a)} \tag{A.16}$$

where $E_{\mathrm{B}}(s,a)$ is the blocking probability in the corresponding M/M/s/s loss system given in Equation (A.6).

We now find the distribution function for the waiting time W by assuming the first-come first-served (FCFS) discipline. Suppose that no servers are idle during a time interval x. Then, the number of service completions during x follows the Poisson distribution with mean $s\mu x$. Let N^- be the number of calls present in the system immediately before an arrival. Hence

$$\begin{aligned} P\{W > x \mid N^- = k\} &= P\{\text{Number of service completions before } x \le k-s\} \\ &= \sum_{j=0}^{k-s} \frac{(s\mu x)^j}{j!}\,\mathrm{e}^{-s\mu x}, \quad x \ge 0. \end{aligned} \tag{A.17}$$

It follows from PASTA again and the theorem of total probability that

$$\begin{aligned} P\{W > x\} &= \sum_{k=s}^{\infty} p_k \cdot P\{W > x \mid N^- = k\} \\ &= \sum_{k=s}^{\infty} p_s \left(\frac{a}{s}\right)^{k-s} \sum_{j=0}^{k-s} \frac{(s\mu x)^j}{j!}\,\mathrm{e}^{-s\mu x} \\ &= p_s\,\mathrm{e}^{-s\mu x} \sum_{j=0}^{\infty} \frac{(s\mu x)^j}{j!} \sum_{k=s+j}^{\infty} \left(\frac{a}{s}\right)^{k-s} \\ &= p_s\,\mathrm{e}^{-s\mu x} \sum_{j=0}^{\infty} \frac{(s\mu x)^j}{j!} \cdot \left(\frac{a}{s}\right)^j \frac{1}{1-a/s} \\ &= \frac{p_s s}{s-a}\,\mathrm{e}^{-s\mu x} \sum_{j=0}^{\infty} \frac{(a\mu x)^j}{j!} \\ &= E_{\mathrm{C}}(s,a)\,\mathrm{e}^{-(s-a)\mu x}, \quad x \ge 0. \end{aligned} \tag{A.18}$$

This result is used in the methodology of required capacity calculation for the packet-switched services in IMT-2000 systems in Section 3.3.2.

Though not used in the methodology, it is interesting to note that the distribution function for the waiting time of those calls that are forced to wait upon arrival is given by

$$\begin{aligned} P\{W > x \mid W > 0\} &= \frac{P\{W > x, W > 0\}}{P\{W > 0\}} = \frac{P\{W > x\}}{P\{W > 0\}} \\ &= \mathrm{e}^{-(s-a)\mu x}, \quad x > 0, \end{aligned} \tag{A.19}$$

which is an exponential distribution with average $1/[(s-a)\mu]$.

We also note that the time average of the number of calls present in the waiting room, denoted by L, at an arbitrary time is given by

$$E[L] = \sum_{k=s+1}^{\infty} (k-s)\, p_k = p_s \sum_{k=s+1}^{\infty} (k-s) \left(\frac{a}{s}\right)^{k-s} = \frac{p_s a s}{(s-a)^2}. \tag{A.20}$$

The waiting time averaged over the calls is given by

$$E[W] = \int_0^{\infty} P\{W > x\}\, \mathrm{d}x = \frac{E_C(s,a)}{(s-a)\mu} = \frac{p_s s}{(s-a)^2 \mu}. \tag{A.21}$$

Hence we can confirm the relation

$$E[L] = \lambda E[W]. \tag{A.22}$$

This is an instance of *Little's law* (Wolff 1989, pp. 235–238) applied to the calls in the waiting room.

Erlang-C formula and the waiting time distribution are derived in Cooper (1991, pp. 90–98), Fujiki and Gambe (1980, p. 75), Gross and Harris (1998, pp. 69–73), Kleinrock (1975, p. 103), and Syski (1986, pp. 238–239).

A.3 Multidimensional Erlang-B Formula

An M/M/s/s loss system with N classes of calls and multiple server occupation is used in the methodology of required capacity calculation for the circuit-switched service categories in IMT-Advanced systems in Section 4.3.4. Before presenting its analysis and computational algorithm which may look formidable at first glance, we study a simple case with two classes of calls and single server occupation. The two-dimensional state transition rate diagram in Figure A.5 helps the reader understand the analysis of the latter system.

A.3.1 Two classes of calls with single server occupation

We first consider an M/M/s/s loss system with two classes of calls. There are s servers and no waiting room. Calls of class 1 and class 2 arrive in independent Poisson processes with rate λ_1 and with rate λ_2, respectively. Each call occupies one server. The service time of each call has exponential distribution with average $1/\mu_1$ for class 1 and with average $1/\mu_2$ for class 2. Any call that arrives when all servers are busy is blocked.

We define the state of the system by a combination of the number of calls of class 1 and the number of calls of class 2 that are present in the system. It is denoted by (j, k), where j and k are the numbers of calls of class 1 and class 2, respectively, present in the system. The state space is finite

$$\Omega := \{(j,k) : j + k \le s,\, j \ge 0,\, k \ge 0\}. \tag{A.23}$$

The number of states is $(s+1)(s+2)/2$. The state follows a two-dimensional birth-and-death process, for which the state transition rate diagram is shown in Figure A.5.

Let $p_{j,k}$ be the probability that there are j calls of class 1 and k calls of class 2 present in the system at an arbitrary time in equilibrium. Referring to Figure A.6(a), the set of *global*

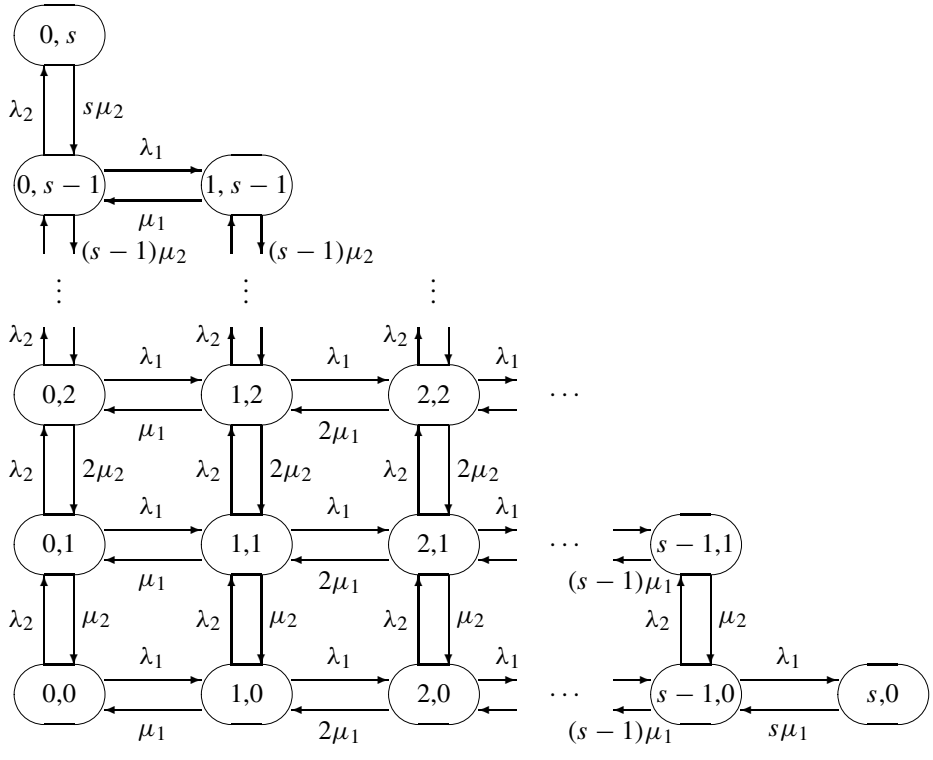
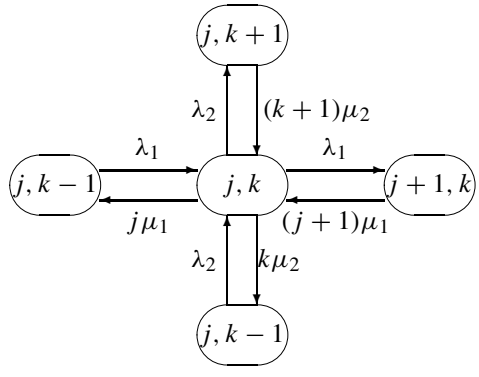


Figure A.5 State transition rate diagram for an M/M/s/s loss system with two classes.

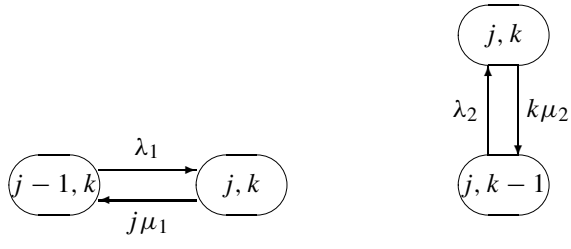
balance equations for $\{p_{j,k};\,(j,k)\in\Omega\}$ is given by

$$\begin{aligned}
\mu_1 p_{1,0}+\mu_2 p_{0,1} &= (\lambda_1+\lambda_2)p_{0,0}\\
\lambda_1 p_{j-1,0}+(j+1)\mu_1 p_{j+1,0}+\mu_2 p_{j,1} &= (\lambda_1+\lambda_2+j\mu_1)p_{j,0}, \quad 1\le j\le s-1\\
\lambda_1 p_{s-1,0} &= s\mu_1 p_{s,0}\\
\lambda_2 p_{0,k-1}+\mu_1 p_{1,k}+(k+1)\mu_2 p_{0,k+1} &= (\lambda_1+\lambda_2+k\mu_2)p_{0,k}, \quad 1\le k\le s-1\\
\lambda_2 p_{0,s-1} &= s\mu_2 p_{0,s} \qquad\qquad (A.24)\\
\lambda_1 p_{j-1,k}+\lambda_2 p_{j,k-1}+(j+1)\mu_1 p_{j+1,k}+(k+1)\mu_2 p_{j,k+1}&\\
= (\lambda_1+\lambda_2+j\mu_1+k\mu_2)p_{j,k}, \quad 1\le j,k\le s-1,\ j+k\le s-1&\\
\lambda_1 p_{j-1,k}+\lambda_2 p_{j,k-1} = (j\mu_1+k\mu_2)p_{j,k}, \quad 1\le j,k\le s-1,\ j+k=s.&
\end{aligned}$$

The number of equations is $(s+1)(s+2)/2$, which is the same as the number of states. One of the equations is redundant as they are homogeneous. Another equation to determine the



(a) Global balance



(b) Local balance

Figure A.6 Global and local balance for an M/M/s/s loss system with two classes.

set $\{p_{j,k};\ (j,k) \in \Omega\}$ uniquely is given by the normalization condition in Equation (A.27) below.

The set of equations in (A.24) is equivalent to the following set of *local balance equations*:

$$\begin{aligned} j\mu_1 p_{j,k} &= \lambda_1 p_{j-1,k}, \quad 1 \le j \le s, 0 \le k \le s \\ k\mu_2 p_{j,k} &= \lambda_2 p_{j,k-1}, \quad 0 \le j \le s, 1 \le k \le s. \end{aligned} \tag{A.25}$$

See Figure A.6(b). These equations are satisfied by the *product-form solution*:

$$p_{j,k} = \frac{1}{G(s)} \frac{(a_1)^j}{j!} \cdot \frac{(a_2)^k}{k!}, \quad (j,k) \in \Omega, \tag{A.26}$$

where $a_1 := \lambda_1/\mu_1$ is the offered traffic of class 1 and $a_2 := \lambda_2/\mu_2$ is the offered traffic of class 2. The constant $G(s)$ can be found from the normalization condition

$$\sum_{(j,k)\in\Omega} p_{j,k} = 1 \tag{A.27}$$

as follows:

$$\begin{aligned} G(s) &= \sum_{(j,k)\in\Omega} \frac{(a_1)^j}{j!}\cdot\frac{(a_2)^k}{k!} = \sum_{j=0}^{s}\sum_{k=0}^{s-j}\frac{(a_1)^j}{j!}\cdot\frac{(a_2)^k}{k!} \\ &= \sum_{i=0}^{s}\sum_{j=0}^{i}\frac{(a_1)^j}{j!}\cdot\frac{(a_2)^{i-j}}{(i-j)!} = \sum_{i=0}^{s}\frac{(a_2)^i}{i!}\sum_{j=0}^{i}\binom{i}{j}\left(\frac{a_1}{a_2}\right)^j \\ &= \sum_{i=0}^{s}\frac{(a_2)^i}{i!}\left(1+\frac{a_1}{a_2}\right)^i = \sum_{i=0}^{s}\frac{(a_1+a_2)^i}{i!}, \end{aligned} \tag{A.28}$$

where we have changed the variable by $j + k = i$ and used the binomial theorem. Hence we obtain the state probability explicitly

$$p_{j,k} = \frac{(a_1)^j/j!\cdot(a_2)^k/k!}{\sum_{i=0}^{s}(a_1+a_2)^i/i!}, \quad (j,k)\in\Omega. \tag{A.29}$$

Let Ω_l be the state space in which the total number of calls in the system is exactly l:

$$\Omega_l := \{(j,k) : j+k=l,\ j\geq 0, k\geq 0\}, \quad 0\leq l\leq s. \tag{A.30}$$

Then we have the probability that there are l calls in the system

$$\begin{aligned} P\{(j,k)\in\Omega_l\} &= \sum_{(j,k)\in\Omega_l} p_{j,k} = \sum_{j=0}^{l} p_{j,l-j} = \frac{1}{G(s)}\sum_{j=0}^{l}\frac{(a_1)^j}{j!}\cdot\frac{(a_2)^{l-j}}{(l-j)!} \\ &= \frac{1}{G(s)}\frac{(a_1+a_2)^l}{l!} = \frac{(a_1+a_2)^l}{l!}\Bigg/\sum_{i=0}^{s}\frac{(a_1+a_2)^i}{i!}, \quad 0\leq l\leq s. \end{aligned} \tag{A.31}$$

This is a truncated Poisson distribution just like Equation (A.5) for an M/M/s/s loss system with a single class of offered traffic $a = a_1 + a_2$.

An arriving call of either class is blocked if all servers are used when it arrives. This happens when the system is in state Ω_s when a call arrives. Due to the PASTA property, the blocking probability is given by

$$P^{\mathrm{B}} = P\{(j,k)\in\Omega_s\} = \frac{(a_1+a_2)^s}{s!}\Bigg/\sum_{i=0}^{s}\frac{(a_1+a_2)^i}{i!}. \tag{A.32}$$

Since

$$\frac{(a_1+a_2)^s}{s!} = \sum_{i=0}^{s}\frac{(a_1+a_2)^i}{i!} - \sum_{i=0}^{s-1}\frac{(a_1+a_2)^i}{i!}, \tag{A.33}$$

the blocking probability can be expressed as

$$P^{\mathrm{B}} = 1 - \frac{G(s-1)}{G(s)}. \tag{A.34}$$

A.3.2 Several classes of calls with multiple server occupation

We now consider an M/M/s/s loss system with N classes of calls and multiple server occupation. There are s servers and no waiting room. Calls of class n arrive in a Poisson process with rate λ_n, independent of the arrival processes of other classes ($1 \leq n \leq N$). It is further assumed that each call of class n occupies m_n servers simultaneously during the service ($1 \leq n \leq N$). Let

$$\boldsymbol{m} \equiv (m_1, m_2, \ldots, m_N).$$

The service time of each call of class n has exponential distribution with average $1/\mu_n$ ($1 \leq n \leq N$). When the service of a call is finished, all servers occupied by the call are released immediately. Any call that arrives when all servers are busy is blocked.

We follow Kaufman (1981) for the analysis and the computational algorithm of the blocking probability. The state of the system is denoted by

$$\boldsymbol{k} \equiv (k_1, k_2, \ldots, k_N) \tag{A.35}$$

where k_n is the number of calls of class n that are present in the system ($1 \leq n \leq N$). The state space is finite

$$\Omega := \{\boldsymbol{k} : \boldsymbol{m} \cdot \boldsymbol{k} \leq s, k_n \geq 0 \ (1 \leq n \leq N)\}, \tag{A.36}$$

where

$$\boldsymbol{m} \cdot \boldsymbol{k} \equiv \sum_{n=1}^{N} m_n k_n. \tag{A.37}$$

In order to write the global balance equations, we introduce the notation

$$\begin{aligned} \boldsymbol{k}_n^+ &\equiv (k_1, \ldots, k_{n-1}, k_n + 1, k_{n+1}, \ldots, k_N) \\ \boldsymbol{k}_n^- &\equiv (k_1, \ldots, k_{n-1}, k_n - 1, k_{n+1}, \ldots, k_N) \end{aligned} \tag{A.38}$$

and

$$\delta_n^+(\boldsymbol{k}) := \begin{cases} 1, & \boldsymbol{k}_n^+ \in \Omega \\ 0, & \text{otherwise;} \end{cases} \qquad \delta_n^-(\boldsymbol{k}) := \begin{cases} 1, & \boldsymbol{k}_n^- \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{A.39}$$

The set of global balance equations is given by

$$\begin{aligned} &\sum_{n=1}^{N} \lambda_n \delta_n^-(\boldsymbol{k}) P(\boldsymbol{k}_n^-) + \sum_{n=1}^{N} (k_n + 1)\mu_n \delta_n^+(\boldsymbol{k}) P(\boldsymbol{k}_n^+) \\ &\quad = \left[\sum_{n=1}^{N} \lambda_n \delta_n^+(\boldsymbol{k}) + \sum_{n=1}^{N} k_n \mu_n \delta_n^-(\boldsymbol{k}) \right] P(\boldsymbol{k}), \quad \boldsymbol{k} \in \Omega. \end{aligned} \tag{A.40}$$

This is equivalent to the following set of local balance equations

$$k_n \mu_n \delta_n^-(\boldsymbol{k}) P(\boldsymbol{k}) = \lambda_n \delta_n^-(\boldsymbol{k}) P(\boldsymbol{k}_n^-), \quad 1 \leq n \leq N, \boldsymbol{k} \in \Omega. \tag{A.41}$$

The solution is given in the product form:

$$P(\boldsymbol{k}) = \frac{1}{G(\Omega)} \prod_{n=1}^{N} \frac{(a_n)^{k_n}}{k_n!}, \quad \boldsymbol{k} \in \Omega \tag{A.42}$$

with

$$G(\Omega) = G(s) = \sum_{\boldsymbol{k}\in\Omega} \prod_{n=1}^{N} \frac{(a_n)^{k_n}}{k_n!} \tag{A.43}$$

where $a_n := \lambda_n/\mu_n$ is the offered traffic of class n ($1 \le n \le N$). It can be easily confirmed by substitution that the solution in (A.42) satisfies Equations (A.41) because

$$P(\boldsymbol{k}_n^-) = \frac{(a_1)^{k_1}}{k_1!} \cdots \frac{(a_{n-1})^{k_{n-1}}}{k_{n-1}!} \cdot \frac{(a_n)^{k_n-1}}{(k_n-1)!} \cdot \frac{(a_{n+1})^{k_{n+1}}}{k_{n+1}!} \cdots \frac{(a_N)^{k_N}}{k_N!}. \tag{A.44}$$

An arriving call of either class is blocked if all servers are used when it arrives. The state space in which an arriving call of class n is blocked is given by

$$B_n := \{\boldsymbol{k} \in \Omega : \boldsymbol{k}_n^+ \notin \Omega\} = \Omega \setminus \{\boldsymbol{k} \in \Omega : 0 \le \boldsymbol{m}\cdot\boldsymbol{k} \le s - m_n\}. \tag{A.45}$$

Therefore, the blocking probability for calls of class n is given by

$$P_n^{\mathrm{B}}(s) = \sum_{\boldsymbol{k}\in B_n} P(\boldsymbol{k}) = \frac{G(B_n)}{G(\Omega)} \tag{A.46}$$

where

$$G(B_n) = \sum_{\boldsymbol{k}\in B_n} \prod_{i=1}^{N} \frac{(a_i)^{k_i}}{k_i!} = G(s) - G(s-m_n). \tag{A.47}$$

Thus we obtain

$$P_n^{\mathrm{B}}(s) = 1 - \frac{G(s-m_n)}{G(s)}. \tag{A.48}$$

We proceed to the computational algorithm. We define

$$q(j) := P\{\boldsymbol{m}\cdot\boldsymbol{k} = j\} = \sum_{\boldsymbol{k}\in\Omega_j} P(\boldsymbol{k}), \quad 0 \le j \le s \tag{A.49}$$

where Ω_j is the state space in which j servers are occupied:

$$\Omega_j := \{\boldsymbol{k} \in \Omega : \boldsymbol{m}\cdot\boldsymbol{k} = j, k_n \ge 0\ (1 \le n \le N)\}, \quad 0 \le j \le s. \tag{A.50}$$

It is clear that the set $\{\Omega_j; 0 \le j \le s\}$ is a partition of Ω:

$$\Omega_i \cap \Omega_j = \emptyset \quad i \ne j; \quad \Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_s. \tag{A.51}$$

It follows that

$$\sum_{j=0}^{s} q(j) = \sum_{j=0}^{s} \sum_{\boldsymbol{k}\in\Omega_j} P(\boldsymbol{k}) = \sum_{\boldsymbol{k}\in\Omega} P(\boldsymbol{k}) = 1. \tag{A.52}$$

Then we have

$$G(j) = G(s) \sum_{l=0}^{j} q(l), \quad 0 \le j \le s. \tag{A.53}$$

Conversely, since $1 = G(0) = G(s)q(0)$ we obtain

$$q(0) = \frac{1}{G(s)}; \quad q(j) = \frac{G(j) - G(j-1)}{G(s)}, \quad 1 \le j \le s. \tag{A.54}$$

The key for the algorithm is the following lemma.

Lemma

$$a_n q(j-m_n) = q(j)E[k_n \mid \boldsymbol{m}\cdot\boldsymbol{k} = j], \quad 1 \le n \le N,\ 0 \le j \le s. \tag{A.55}$$

Proof. The local balance equations in (A.41) can be written as

$$a_n \delta_n(\boldsymbol{k}) P(\boldsymbol{k}_n^-) = k_n P(\boldsymbol{k}), \quad 1 \le n \le N \tag{A.56}$$

where

$$\delta_n(\boldsymbol{k}) := \begin{cases} 1, & k_n \ge 1 \\ 0, & \text{otherwise.} \end{cases} \tag{A.57}$$

Summing both sides of (A.56) over Ω_j, we obtain

$$a_n \sum_{\boldsymbol{k}\in\Omega_j} \delta_n(\boldsymbol{k}) P(\boldsymbol{k}_n^-) = \sum_{\boldsymbol{k}\in\Omega_j} k_n P(\boldsymbol{k}). \tag{A.58}$$

We calculate each side of this equation. First, by noting that

$$P(\boldsymbol{k} \mid \boldsymbol{m}\cdot\boldsymbol{k} = j) = \begin{cases} P(\boldsymbol{k})/q(j), & \boldsymbol{k} \in \Omega_j \\ 0, & \text{otherwise} \end{cases} \tag{A.59}$$

the right-hand side of (A.58) becomes

$$\begin{aligned} \sum_{\boldsymbol{k}\in\Omega_j} k_n P(\boldsymbol{k}) &= \sum_{\boldsymbol{k}\in\Omega_j} k_n \frac{P(\boldsymbol{k})}{q(j)} \cdot q(j) \\ &= q(j) \sum_{\boldsymbol{k}\in\Omega_j} k_n P(\boldsymbol{k} \mid \boldsymbol{m}\cdot\boldsymbol{k} = j) \\ &= q(j)E[k_n \mid \boldsymbol{m}\cdot\boldsymbol{k} = j]. \end{aligned} \tag{A.60}$$

Next, on the left-hand side of (A.58), we have

$$\sum_{\boldsymbol{k}\in\Omega_j} \delta_n(\boldsymbol{k}) P(\boldsymbol{k}_n^-) = \sum_{\boldsymbol{k}\in\Omega_j \cap \{k_n \ge 1\}} P(\boldsymbol{k}_n^-). \tag{A.61}$$

However, since

$$\boldsymbol{m}\cdot\boldsymbol{k} = \sum_{l=1}^{N} m_l k_l = \sum_{l\ne n}^{N} m_l k_l + m_n(k_n - 1) + m_n = \boldsymbol{m}\cdot\boldsymbol{k}_n^- + m_n, \tag{A.62}$$

it follows that

$$\begin{aligned} \Omega_j \cap \{k_n \ge 1\} &= \{\boldsymbol{k} : \boldsymbol{m}\cdot\boldsymbol{k} = j, k_n \ge 1, k_l \ge 0\ (l \ne n)\} \\ &= \{\boldsymbol{k}_n^- : \boldsymbol{m}\cdot\boldsymbol{k}_n^- = j - m_n, (\boldsymbol{k}_n^-)_l \ge 0\ (1 \le l \le N)\}. \end{aligned} \tag{A.63}$$

Hence we obtain

$$\sum_{\boldsymbol{k}\in\Omega_j} \delta_n(\boldsymbol{k}) P(\boldsymbol{k}_n^-) = q(j - m_n). \tag{A.64}$$

q. e. d.

This lemma leads to the following theorem.

Theorem

$$\sum_{n=1}^{N} a_n m_n q(j - m_n) = jq(j), \quad 1 \le j \le s \tag{A.65}$$

where $q(j) = 0$ *for* $j < 0$.

Proof. Multiplying both sides of (A.55) by m_n and summing over n yields

$$\begin{aligned}\sum_{n=1}^{N} a_n m_n q(j - m_n) &= q(j) \sum_{n=1}^{N} m_n E[k_n \mid \boldsymbol{m} \cdot \boldsymbol{k} = j] \\ &= q(j) E\left[\sum_{n=1}^{N} m_n k_n \mid \boldsymbol{m} \cdot \boldsymbol{k} = j\right] \\ &= jq(j). \end{aligned} \tag{A.66}$$

q. e. d.

Finally we obtain the following recursive relation for $\{G(j)\ 0 \le j \le s\}$.

Corollary

$$jG(j) = \sum_{l=0}^{j-1} G(l) + \sum_{n=1}^{N} a_n m_n G(j - m_n), \quad 1 \le j \le s. \tag{A.67}$$

Proof. From (A.65), we have for $1 \le l \le s$

$$\begin{aligned} G(s) \sum_{n=1}^{N} a_n m_n q(l - m_n) &= l \cdot G(s) q(l) = l[G(l) - G(l-1)] \\ &= lG(l) - (l-1)G(l-1) - G(l-1). \end{aligned} \tag{A.68}$$

Summing the both sides over $l = 1, 2, \ldots, j$ we obtain

$$\sum_{n=1}^{N} a_n m_n G(j - m_n) = jG(j) - \sum_{l=0}^{j-1} G(l). \tag{A.69}$$

q. e. d.

The recursive formula in Equation (A.65) is usually called *Kaufman–Roberts algorithm* (Kaufman 1981; Roberts 1981), although it was first derived by Fortet and Grandjean (1964). Equation (A.67) is shown by Takagi *et al.* (2006). It is used in the methodology of required capacity calculation for the circuit-switched service categories in IMT-Advanced systems in Section 4.3.4.

A.4 M/G/1 Nonpreemptive Priority Queue

We consider a single server queue with an infinite waiting room. There are N classes of calls indexed $n = 1, 2, \ldots, N$. Calls of class n arrive in a Poisson process at rate λ_n. The service times of calls of class n have general distribution with average b_n and second moment $b_n^{(2)}$. We assume that the classes of calls are *priority classes* with the descending order of indices. That is, class n has priority over class i if and only if $n < i$. Class 1 is the highest priority class and class N the lowest. When the server becomes available, it selects for service one call of which priority is the highest among those calls present in the waiting room. We assume the FCFS discipline for the order of service within the same class of calls. The service is *nonpreemptive* in the sense that once the service to a call is started it is not disrupted until the completion (even if any calls of higher priority arrive during the service).

We calculate the average waiting time for calls of class n, denoted by W_n for $1 \leq n \leq N$ (Kleinrock 1976, pp. 119–121). To do so, we focus on a newly arriving call of class n which is referred to as a 'tagged' call. The waiting time of this tagged call consists of the following components:

- delay due to the call in service when our tagged call arrives;
- delay due to those calls found in the waiting room that receive service before our tagged call;
- delay due to those calls which arrive at the system after our tagged call but receive service before the tagged call.

Consequently, the average waiting time for our tagged call can be written as

$$W_n = W_0 + \sum_{i=1}^{N} b_i (L_{i,n} + M_{i,n}), \quad 1 \leq n \leq N, \tag{A.70}$$

where

$$W_0 := \text{average delay to our tagged call due to the call found in service when our tagged call arrives} \tag{A.71}$$

$$L_{i,n} := \text{average number of calls of class } i \text{ found in the waiting room by our tagged call of class } n \text{ and which receive service before the tagged call} \tag{A.72}$$

$$M_{i,n} := \text{average number of calls of class } i \text{ that arrive at the system while our tagged call of class } n \text{ is in the waiting room and which receive service before the tagged call.} \tag{A.73}$$

We first evaluate W_0. The delay to our tagged call due to a call of class i found in service by the tagged call is the remaining service time of the call of class i when the tagged call arrives. The average remaining service time of a call of class i is derived as follows (Kleinrock 1975, pp. 169–173).

Let $f_i(x)$ be the probability density function (pdf) for the service time of a call of class i. It is important to note that the pdf $\hat{f}_i(x)$ for the service time $\hat{X}_i$ of a call of class i during which a call arrives is generally different from $f_i(x)$. This is because long service times occupy larger segments of the time axis than short service times do, and therefore it is more likely that an arbitrary call arrives during a long service time. The probability that a call arrives during the service time of a call of class i with duration x is proportional to the duration x as well as to the probability density $f_i(x)$. Thus

$$\hat{f}_i(x) = cxf_i(x), \tag{A.74}$$

where c is a constant. We can obtain $c = 1/b_i$ from the normalization condition as

$$1 = \int_0^\infty \hat{f}_i(x)\,\mathrm{d}x = c\int_0^\infty xf_i(x)\,\mathrm{d}x = cb_i. \tag{A.75}$$

Hence we obtain

$$\hat{f}_i(x) = \frac{xf_i(x)}{b_i}. \tag{A.76}$$

Let us now find the pdf $g_i(y)$ for the remaining service time Y_i of a call of class i. Given that our tagged call arrives during the service time of a call of class i with duration x, the arrival point is uniformly distributed over an interval $[0, x]$. Then we have

$$P\{Y_i \le y\} = \frac{y}{x}, \quad 0 \le y \le x. \tag{A.77}$$

Hence the joint density of $\hat{X}_i$ and Y_i is given by

$$P\{y < Y_i \le y + \mathrm{d}y, x < \hat{X}_i \le x + \mathrm{d}x\} = \frac{\mathrm{d}y}{x} \cdot \frac{xf_i(x)\,\mathrm{d}x}{b_i} = \frac{f_i(x)\,\mathrm{d}y\,\mathrm{d}x}{b_i}, \quad 0 \le y \le x. \tag{A.78}$$

Integrating over x, we obtain the pdf for Y_i as

$$g_i(y) = \int_{x=y}^{x=\infty} \frac{f_i(x)\,\mathrm{d}x}{b_i} = \frac{1}{b_i}\int_y^\infty f_i(x)\,\mathrm{d}x = \frac{1 - F_i(y)}{b_i} \tag{A.79}$$

where $F_i(y) := \int_0^y f_i(x)\,\mathrm{d}x$ is the distribution function of X_i. We then obtain

$$E[Y_i] = \int_0^\infty yg_i(y)\,\mathrm{d}y = \frac{1}{b_i}\int_0^\infty y[1 - F_i(y)]\,\mathrm{d}y = \frac{b_i^{(2)}}{2b_i} \tag{A.80}$$

where the last result can be calculated by the method of partial integration.

With the above preparation, we can proceed to obtain W_0. To do so, we note that

$$\rho_i := \lambda_i b_i \tag{A.81}$$

is the fraction of time that the server is occupied by calls of class i. From the PASTA property, this is also the probability that our tagged call finds a call of class i in service. The average time until the completion of such service is given by $b_i^{(2)}/(2b_i)$. Hence we obtain

$$W_0 = \left(1 - \sum_{i=1}^N \rho_i\right) \cdot 0 + \sum_{i=1}^N \rho_i \cdot \frac{b_i^{(2)}}{2b_i} = \frac{1}{2}\sum_{i=1}^N \lambda_i b_i^{(2)}, \tag{A.82}$$

where the factor $1 - \sum_{i=1}^{N} \rho_i$ in the middle expression is the probability that there are no calls in service when our tagged call arrives.

We next evaluate $L_{i,n}$. It is clear that those calls of classes $n+1, n+2, \ldots, N$ (lower priority) found in the waiting room by our tagged call do not delay the tagged call. Hence

$$L_{i,n} = 0, \quad n+1 \le i \le N. \tag{A.83}$$

On the other hand, those calls of classes $1, 2, \ldots, n$ (higher or equal priority) which are present in the waiting room upon the arrival of our tagged call must certainly be served before the tagged call. It follows from Little's law that the average number of such calls of class i present in the waiting room at an arbitrary time is given by $\lambda_i W_i$. From PASTA, this is also true at the arrival time of the tagged call. Thus we have

$$L_{i,n} = \lambda_i W_i, \quad 1 \le i \le n. \tag{A.84}$$

We finally evaluate $M_{i,n}$. Not only those calls of classes $n+1, n+2, \ldots, N$ (lower priority) but also the calls of class n (the same priority) that arrive while our tagged call is in the waiting room do not delay the tagged call. The latter is due to the assumption that calls within the same class are served according to the FCFS discipline. Hence

$$M_{i,n} = 0, \quad n \le i \le N. \tag{A.85}$$

Those calls of classes $1, 2, \ldots, n-1$ (higher priority) that arrive while our tagged call is in the waiting room will also be served before the tagged call. The average time that the tagged call spends in the waiting room is W_n. Since the arrival rate of calls of class i is λ_i and since the arrival process of each class is independent of the number of calls in the waiting room, there will be on average $\lambda_i W_n$ arrivals of calls of class i while our tagged call is in the waiting room. Thus we have

$$M_{i,n} = \lambda_i W_n, \quad 1 \le i \le n-1. \tag{A.86}$$

Note that this is *not* from Little's law.

Substituting Equations (A.82) to (A.86) into Equation (A.70), we obtain

$$\begin{aligned} W_n &= W_0 + \sum_{i=1}^{n} b_i \lambda_i W_i + \sum_{i=1}^{n-1} b_i \lambda_i W_n \\ &= W_0 + \sum_{i=1}^{n} \rho_i W_i + W_n \sum_{i=1}^{n-1} \rho_i \\ &= W_0 + \sum_{i=1}^{n-1} \rho_i W_i + W_n \sum_{i=1}^{n} \rho_i, \quad 1 \le n \le N, \end{aligned} \tag{A.87}$$

which can be written as

$$W_n = \frac{W_0 + \sum_{i=1}^{n-1} \rho_i W_i}{1 - \sum_{i=1}^{n} \rho_i}, \quad 1 \le n \le N, \tag{A.88}$$

where the null sum $\sum_{i=1}^{0}$ is zero. This triangular set of simultaneous equations can be solved recursively by starting with W_1, then calculating W_2, and so on to obtain

$$W_n = \frac{W_0}{(1-\sum_{i=1}^{n-1}\rho_i)(1-\sum_{i=1}^{n}\rho_i)} = \frac{\sum_{i=1}^{N}\lambda_i b_i^{(2)}}{2(1-\sum_{i=1}^{n-1}\lambda_i b_i)(1-\sum_{i=1}^{n}\lambda_i b_i)}, \quad 1 \le n \le N. \tag{A.89}$$

This formula was first derived by Cobham (1954). It is used in the methodology of required capacity calculation for the packet-switched service categories in IMT-Advanced systems in Section 4.3.5.

We may confirm that the inequality

$$W_{n-1} < W_n, \quad 1 \le n \le N, \tag{A.90}$$

which demonstrates the effect of prioritized handling. It is also interesting to note the relation

$$\sum_{n=1}^{N}\rho_n W_n = \frac{\rho W_0}{1-\rho} \tag{A.91}$$

where $\rho := \sum_{n=1}^{N}\rho_n$. This can be easily seen from

$$\frac{\rho_n}{(1-\sum_{i=1}^{n-1}\rho_i)(1-\sum_{i=1}^{n}\rho_i)} = \frac{1}{1-\sum_{i=1}^{n}\rho_i} - \frac{1}{1-\sum_{i=1}^{n-1}\rho_i}, \quad 1 \le n \le N. \tag{A.92}$$

The relation in (A.91) holds for any M/G/1 queue with multiple classes that has nonpreemptive work-conserving service discipline (Kleinrock 1976, pp. 113–117). This relationship is called *Kleinrock's conservation law*.