From Introduction to Queueing Theory
by B.V. Gnedenko and I.N. Kovalenko
1968

Since in our case $\lambda\tau=\rho=1$, we have

$$\tau s=\frac{\mu_2}{2}(s+\lambda\varepsilon)^2+o((s+\lambda\varepsilon)^2),$$

whence

$$\varepsilon=\tau\sqrt{\frac{2\tau s}{\mu_2}}\,[1+o(1)].$$

Hence the decomposition

$$\varphi(s)=\sqrt{\frac{\mu_2}{2\tau s}}+o\left(\frac{1}{\sqrt{s}}\right) \qquad (s\to 0),$$

and by the Tauberian theorem

$$F(t,+0)=\sqrt{\frac{\mu_2}{2\pi\tau t}}+o\left(\frac{1}{\sqrt{t}}\right) \qquad (t\to\infty). \tag{7}$$

Expression (7) may be sharpened in a similar manner, by using decompositions of higher order.

## 4.4. THE HEAVILY-LOADED SINGLE-SERVER SYSTEM

### 1. Load on server

When investigating systems with waiting, the quantity $\rho=\lambda\tau$ plays a special role; here $\lambda$ is the parameter of the incoming stream and $\tau$ is the mean waiting time. The results of Sec. 4.1 and Sec. 4.2 indicate that for $\rho<1$ the stochastic processes connected with the functioning of the serving system have an ergodic distribution. On the other hand when $\rho\geqslant 1$ the serving system is characterized by the fact that, with probability arbitrarily close to 1, a customer arriving sufficiently late has an arbitrarily long waiting time. It was also proved that the proportion of the time in which the server is loaded is equal to $\rho$ when $\rho\leqslant 1$. $\rho$ is therefore called the load on the server.

### 2. Limit theorems

It is interesting to study the case in which the load is not exactly 1 but close to 1. Namely, let us assume that $\rho=1-\varepsilon$, $\varepsilon>0$ and study the limiting situation when $\varepsilon\to 0$. The mathematical expectation will increase to infinity since it is of the order of magnitude $\frac{1}{\varepsilon}$. Instead of the waiting time $\gamma$ and the number of customers in the system $\nu$ we shall, therefore, consider the normalized quantities

$$\bar\gamma=\varepsilon\gamma, \quad \bar\nu=\varepsilon\nu.$$

*Theorem.* *When $\rho\to 1$ and the service time has finite variance $\sigma^2$, the distribution function of the random variable $\bar\gamma$ converges to the exponential distribution function*

$$P\{\bar\gamma>x\}\to e^{-\frac{2\tau x}{\sigma^2+\tau^2}}. \tag{1}$$

P r o o f. The ergodic distribution of the waiting time has, as shown in Sec. 4.1, the following Laplace-Stieltjes transform:

$$\Phi(s)=\frac{1-\rho}{1-\lambda\,\frac{1-h(s)}{s}}.$$

The characteristic function of this distribution, equal to the Laplace-Stieltjes transform for pure imaginary values of the argument

$$\varphi(t)=\Phi(-it)$$

is

$$\varphi(t)=\frac{1-\rho}{1-i\lambda\,\frac{1-\psi(t)}{t}}\qquad(\psi(t)=h(-it)). \tag{2}$$

It is well-known that if a sequence of characteristic functions converges to some characteristic function, the corresponding sequence of distribution functions converges to the distribution function corresponding to the limiting characteristic function. To prove the theorem it is, therefore, sufficient to show that as $\rho\to 1$ the characteristic function of the random variable $\bar{\gamma}$ converges to

$$\frac{1}{1-\frac{it}{2}\left(\frac{\sigma^2+\tau^2}{\tau}\right)}, \tag{3}$$

which is the characteristic function of an exponential distribution with mathematical expectation $\frac{\sigma^2+\tau^2}{2\tau}$. We recall that multiplication of a random variable by a constant corresponds to the multiplication of the argument of the characteristic function by the same constant. Thus, if $\bar{\varphi}(t)$ is the characteristic function of the random variable $\bar{\gamma}$, then

$$\bar{\varphi}(t)=\varphi(\varepsilon t)=\frac{1-\rho}{1-i\lambda\,\frac{1-\psi(\varepsilon t)}{\varepsilon t}}=\frac{\varepsilon}{1-i\lambda\,\frac{1-\psi(\varepsilon t)}{\varepsilon t}}.$$

Since by assumption the service time has finite variance, the characteristic function of this random variable can be represented in a neighborhood of zero by Taylor's formula

$$\psi(t)=1+i\tau t-\frac{t^2}{2}(\sigma^2+\tau^2)+o(t^2).$$

Substituting this in the preceding expression and letting $\varepsilon\to 0$ we get (3). The theorem is proved.

An analogous result holds for the distribution of the number of customers in the system at an arbitrary instant under steady-state conditions.

*T h e o r e m. When $\rho\to 1$ and the service time has finite variance $\sigma^2$, the distribution function of the random variable $\bar{\nu}$ converges to the exponential distribution function*

$$P\{\bar{\nu}>x\}\underset{\rho\to 1}{\longrightarrow}e^{-\frac{2\tau^2 x}{\sigma^2+\tau^2}}.$$

P r o o f. We shall use the Pollaczek-Khinchin formula for the generating function of the distribution of the number of customers in the system:

$$\sum_{n=0}^{\infty}p_n z^n=\frac{(1-\rho)(1-z)\,h(\lambda(1-z))}{h(\lambda(1-z))-z}.$$

If $\pi(z)$ is the generating function of some random variable and $\varphi(t)$ the characteristic function, we have

$$\varphi(t)=\pi(e^{it}).$$

Hence the characteristic function of the random variable $\nu$ is

$$\frac{(1-\rho)(1-e^{it})h(\lambda(1-e^{it}))}{h(\lambda(1-e^{it}))-e^{it}}.$$

If $\overline{l}(t)$ denotes the characteristic function of the random variable $\overline{\nu}$, we get

$$\overline{l}(t)=\frac{\varepsilon(1-e^{it\varepsilon})h[\lambda(1-e^{it\varepsilon})]}{h[\lambda(1-e^{it\varepsilon})]-e^{it\varepsilon}}.$$

Since by assumption the service time has finite variance, there exists a Taylor expansion for any finite $t$ and $\varepsilon \to 0$:

$$h(\lambda(1-e^{it\varepsilon}))=1+it\varepsilon\lambda\tau-\frac{t^2\varepsilon^2\lambda}{2}(\lambda\mu_2+\tau)+o(\varepsilon^2).$$

Substituting in the preceding formula and letting $\varepsilon \to 0$, we get

$$\lim_{\varepsilon\to 0}\overline{l}(t)=\frac{1}{1-it\frac{\sigma^2+\tau^2}{2\tau^2}},$$

as was to be proved.

Remark. In Sec. 1.2 we proved that when the service time is exponentially distributed, the distribution law of the waiting time in the stationary case is an exponent with a jump at zero, while the probabilities are geometrically distributed. The limit theorems just proved indicate that when the server is heavily loaded, i.e., when $1-\rho$ is small, the limiting distributions of the waiting time and the number of customers in the system have a similar simple structure.

Theorems of this type, which characterize the asymptotic behavior of the serving process under a load approaching the critical, are valid under assumptions considerably weaker than those made in this section. We shall return to this problem in connection with more complex systems.

## 4.5. SINGLE-SERVER SYSTEM WITH WAITING WITH FAILURE AND RENEWAL OF THE SERVER

### 1. Possible formulations of the problems

Extending the schemes for queueing processes discussed in the foregoing sections to cases in which the server fails and requires renewal (repair) is of great practical importance. It is a priori obvious that, while the waiting time and the queue length may have ergodic distributions for a given intensity of the incoming stream, the queue will in fact tend to infinity if one takes into account time expended on repair due to systematic failures. We must therefore establish conditions for ergodicity and find the different characteristics of service in the case of "unreliable" servers.

From the practical viewpoint we are interested in different mathematical formulations of failure and renewal of the server, and in the principles of