System Analysis and Optimization

Prof.　Li Li

Mobile: 18916087269

E-mail: lili@tongji.edu.cn

Office: E&I-Building, 611

Department of Control Science and Engineering

College of Electronics and Information Engineering

Tongji University

Class Introduction

- **Time & Venue: Tuesday (3,4), C211**
- **Teaching Materials**
  - **Manufacturing Systems Modeling, Analysis and Optimization (PPT)**
  - **Manufacturing Systems Modeling and Analysis. Guy L. Curry, Richard M. Feldman (Eds.), Springer, 2009**
  - **Handbook of Memetic Algorithms. Ferrante Neri, Carlos Cotta, and Pablo Moscato (Eds.), Springer, 2011**
  - **Production Planning and Control for Semiconductor Wafer Fabrication Facilities-Modeling, Analysis and Systems. Lars Monch, John W.Fowler and Scott J.Mason, Springer, 2013**
- **Grading Procedures: Attendance (20%) + Final Exam (80%)**

Main Contents

Part 1: Manufacturing systems modeling and analysis

- **Performance measures, Introduction to factory models**

Part 2: Manufacturing systems optimization

- **Basic Concepts, Optimal approaches, Heuristic methods, Descriptive models**

Part 3: Case Study - Semiconductor Manufacturing System Modeling, Analysis and Optimization

- **Introduction to semiconductor manufacturing system, Release control, Dispatching methods, State of the Practice and Future Needs for Production Planning and Control Systems**

- Intel automation system (video)

- **Example: Semiconductor Wafer Fabrication Facility (Fab)**
  - A semiconductor chip is a highly miniaturized, integrated electronic circuit consisting of thousands of components.
  - The whole manufacturing process may require up to 700 single process steps and up to 3 months to produce.

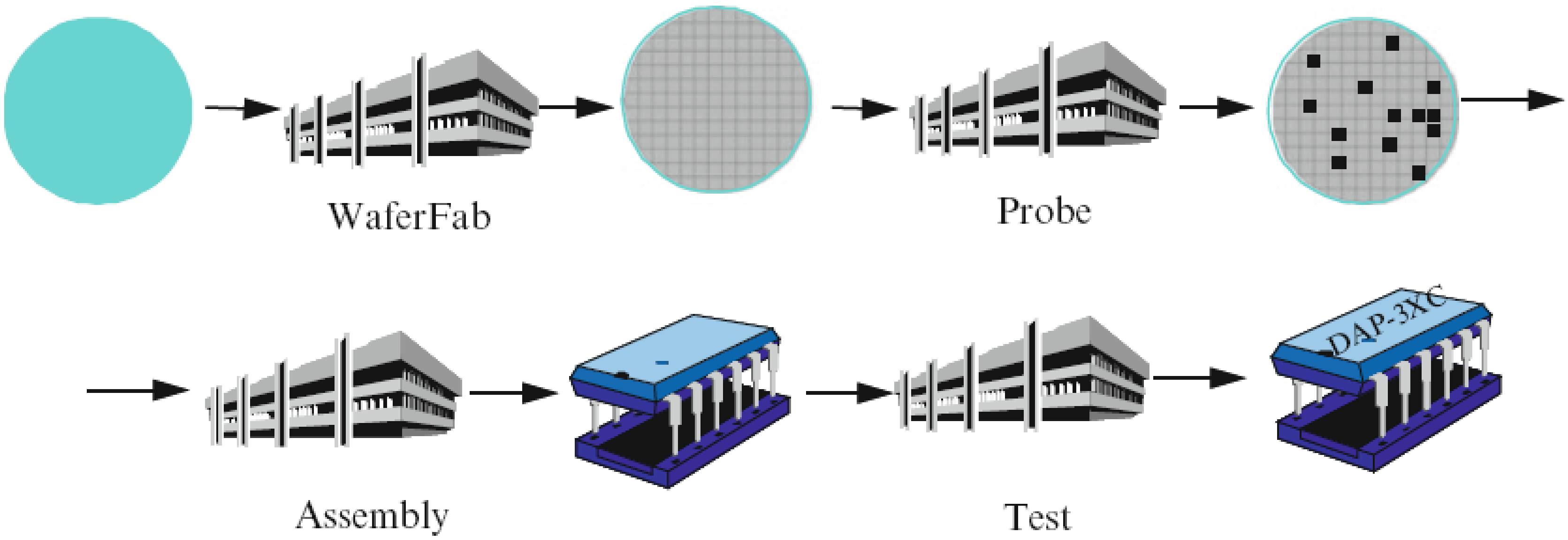


Fig1.1 Stages of semiconductor manufacturing

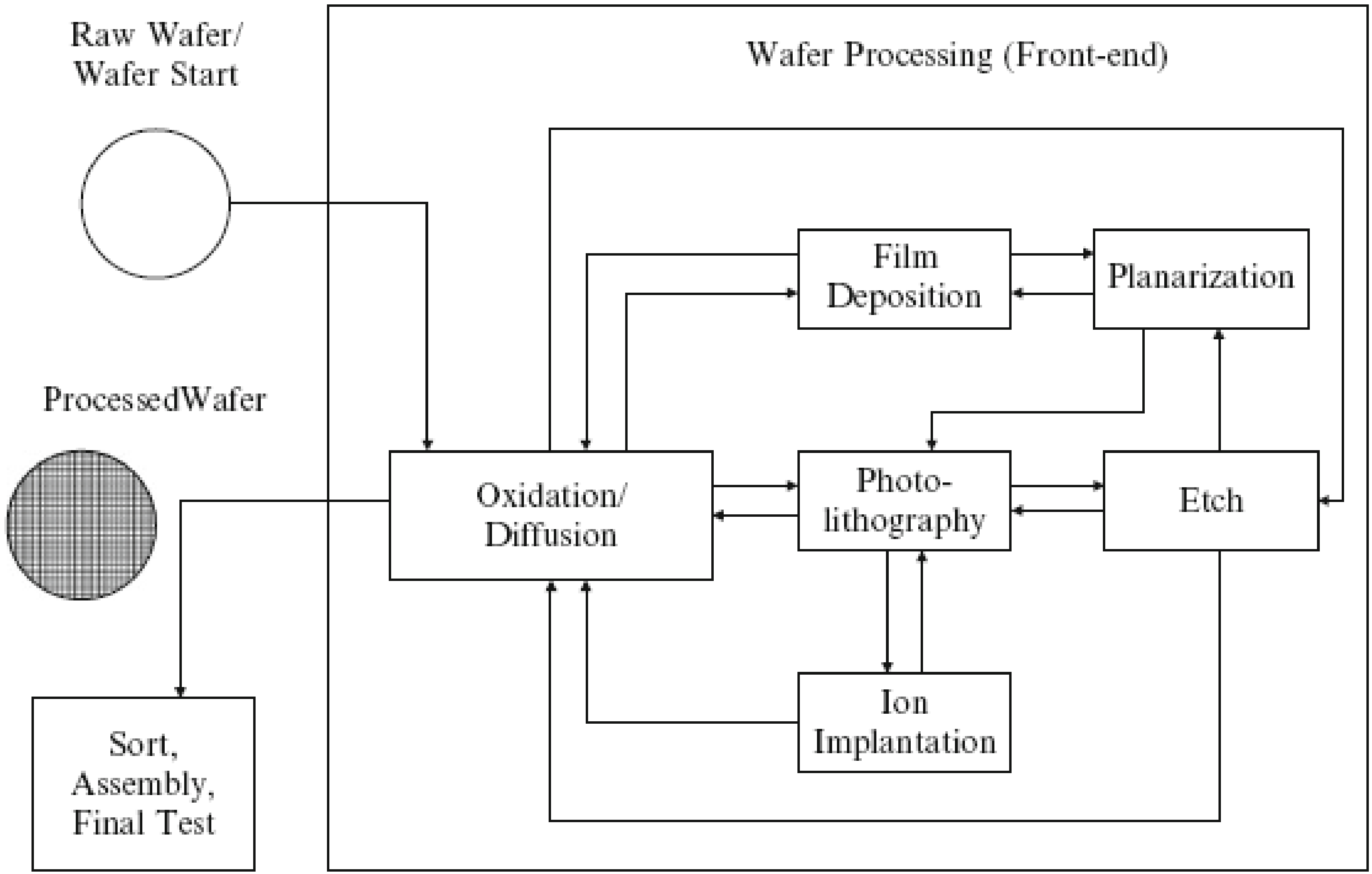


Fig1.2 Operations in a wafer fab

Chapter 1 Introduction to Factory Models

- Definitions
- Performance measures
- Single workstation factory models
- Processing time variability
- Multiple-stage single-product factory models
- Multiple-product factory models

1.1 Definitions

- **Definition 1.1: A *job* is a physical entity that must be processed through the various processing steps or may be an order to begin the processing of raw material into a newly manufactured entity.**
- **Definition 1.2: A *factory* consists of several machines grouped together by type (called workstations) and a series of jobs that are to be produced on these machines.**

-The workflow of a job moving through the factory: waiting in line at a machine (workstation) until its turn for processing → being processed on the machine → proceeding to the next machine location to repeat the sequence until all required operations have been completed.

1.1 Definitions

- **Definition 1.3. A *workstation* (or machine group) is a collection of one or more identical machines or resources.**

  -In a general manufacturing context, workstations are sometimes made up of several different machine types called cells where these machines are gathered together for the purpose of performing several distinct processing steps at one physical location. In order to model a cell type workstation, one would need to combine several single-machine workstations together.

- **Definition 1.4. A *processing step* for a job consists of a specific machine or workstation and the processing time (possibly processing time distribution) for the step.**

1.1 Definitions

- **Definition 1.5. The sequence of processing steps for a job is called its *routing*. Jobs with identical routings are said to be of the same *job type*; thus, different job types are jobs with different routings.**

  -The characteristics of all the job routings determine the organization of a manufacturing facility that is used to produce these jobs.

  - **A unique routing: an *assembly line* given a high enough throughput rate**
  - **A few routings (a low diversity of job types) with each routing visiting a workstation at most one time: a *flow shop***
  - **A large numbers of different job routings (a high diversity of job types) so that jobs visit workstations with no apparent structure: a *job shop***
  - **A given workstation could be visited in several processing steps with the same job routing: a *re-entrant flow***

1.1 Definitions

- **Definition 1.6: *Cycle time* is the time that a job spends within a system. The average cycle time is denoted by CT.**
  - ***CT_s* : the average factory cycle time, i.e., the average time that a job spends with the factory, either being processed at a workstation or waiting in a workstation queue.**
  - ***CT*(*i*): the average cycle time jobs spend being processed by workstation *i* (the *ith* grouping of identical machines) plus the average time spend in the queue (or buffer).**

$$CT(i) = CT_q(i) + T_s(i)$$

-$CT_q(i)$ denotes the average time a job spends in the queue in front of the workstation

-$T_s(i)$ denotes the service time (or processing time) at workstation *i*.

1.1 Definitions

- **Definition 1.7.** ***Work-in-process*** **is the number of jobs within a system that are either undergoing processing or waiting in a queue for processing. The average work-in-process is denoted by** ***WIP*****.**
- **Definition 1.8. The** ***throughput rate*** **for a system is the number of completed jobs leaving the system per unit of time. The throughput rate averaged over many jobs is denoted by** ***th*****.**
  - ✓ **For most of the systems that we will consider, the long-run throughput rate of the system must be equal to the input rate of jobs.**
  - ✓ **Given that the throughput rate is known and there is enough capacity to satisfy the long term average demand, the higher the factory capacity relative to the needs, the faster jobs are completed.**

1.1 Definitions

- **Definition 1.9. The x-factor for a factory is the ratio of CT_s to the average total processing time per job.**
- **Diagram used to illustrate the nature of a modeled system will omit the system level structure and emphasize the internal structure of the model itself. The level of detail generally needed in diagrams will include workstations and job flow within the factory.**

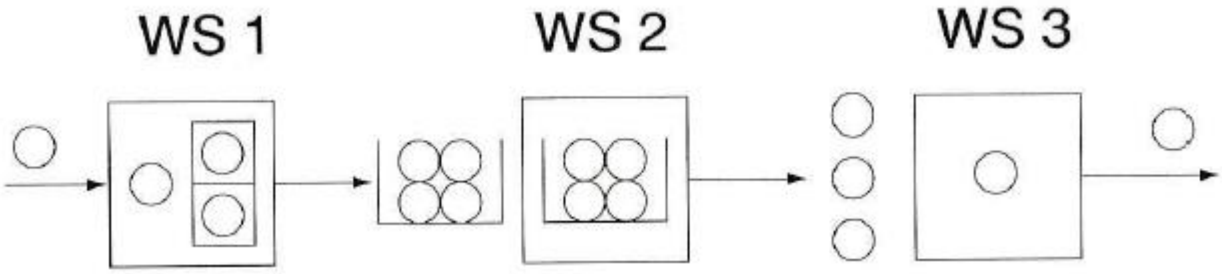
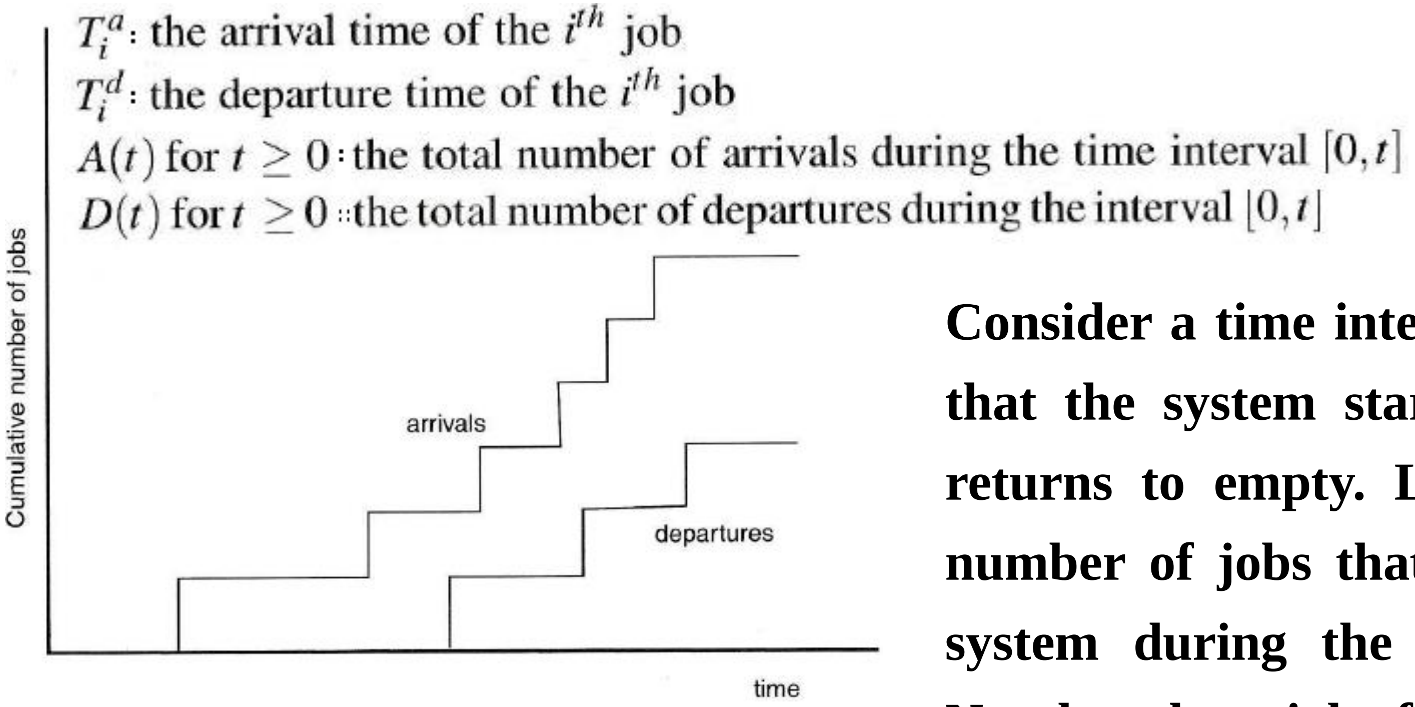


Fig.1.1. Detailed diagram depicting the two machines in Workstation 1, a batch processing operation at Workstation 2, and individual processing on a single machine at Workstation 3

1.2 Performance Measures

- **Measuring *CT* and *WIP*: record the number of arrivals and departures to and from the system**



T_i^a: the arrival time of the i^{th} job

T_i^d: the departure time of the i^{th} job

$A(t)$ for $t \geq 0$: the total number of arrivals during the time interval $[0,t]$

$D(t)$ for $t \geq 0$ ::the total number of departures during the interval $[0,t]$

Fig.1.2. Arrival A(·) and departure D(·) functions for a system in which arrivals and departures occur one at a time.

Consider a time interval (*a*,*b*) such that the system starts empty and returns to empty. Let N_{ab} be the number of jobs that arrive to the system during the interval (*a*,*b*). Number these jobs for 1 to *N*, with index *i* representing specific jobs.

1.2 Performance Measures

- **The average cycle time, *CT*(*a*,*b*), for jobs during this interval is given by**

$$CT(a,b) = \frac{1}{N_{ab}} \sum_{i=1}^{N_{ab}} (T_i^d - T_i^a) .$$

  - **Note: the area (AB) between the curves *A*(*t*) and *D*(*t*) for *a*<*t*<*b* is merely the summation given in the above equation.**

- **The time-averaged number of jobs waiting in the system during the time interval (*a*,*b*) is given by**

$$WIP(a,b) = \frac{1}{b-a} \int_a^b (A(t) - D(t))\mathrm{d}t .$$

$$WIP(a,b) = \frac{1}{b-a} AB \quad \text{and} \quad CT(a,b) = \frac{1}{N_{ab}} AB .$$

$$WIP(a,b) = \frac{N}{b-a} CT(a,b) .$$

1.2 Performance Measures

- **Conclusion: the mean number of jobs arriving to the system per unit time, normally denoted as *λ*, is $N_{ab}/(b-a)$.**

$$WIP(a,b) = \lambda CT(a,b)\,.$$

  - **This result is valid for any interval that starts with an empty system and ends with an empty system.**
  - **The relationship is the limiting behavior result, or long run average result, for stationary queuing systems, and is known as Little's Law.**
  - **The result holds for individual workstations as well as the system as a whole.**

1.2 Performance Measures

- **Property 1.1. Little's Law.** ***For a system that satisfies steady-state conditions, the following equation holds***

$$\boldsymbol{WIP=\lambda\times CT}$$

where WIP is the long-run average number of jobs in the system, CT is the long-run average cycle time and λ is the long-run input rate of jobs to the sever.

- **Since the average input rate is usually equal to the average throughput rate, Little's Law can also be written as *WIP=th×CT*.**

- It should be stressed that the limiting behavior generally estimates mean values and the actual underlying random variables for the systems can be quite variable. It is often desired that analytical models of these systems describe the steady state probability distribution.

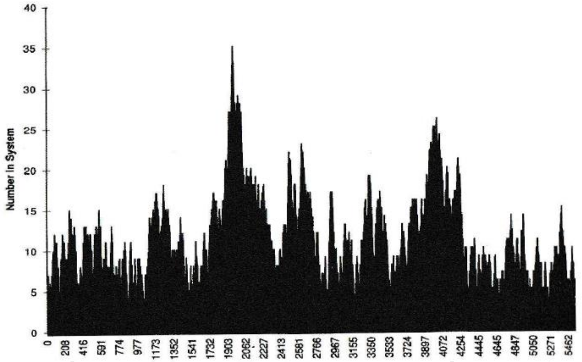


Fig.1.3. A representation of the number of jobs in a simulated factory

- **Example 1.1. Consider a factory that makes only one type of product. The processing requirements for this product consists of four processing steps that must be performed in sequence. Each processing operation is performed on a separate machine. These machines can process only one unit of the product at a time (called a job). The processing times for the four operations are constant. These processing times are 1, 2,1 and 1 hour(s) on each of the four machines, respectively. This idealized factory has no machine downtimes, no product unit losses due to faulty production, and operates continuously. The factory is operated using a constant number of jobs in process (i.e., WIP_s (t) is constant for all *t*). When a job has completed its four processing steps, it is immediately removed from the factory and a new job is started at Machine 1 to keep the total factory WIP_s at the specified level. This process is depicted in Fig.1.4. This factory is running smoothly at the current time. Management has set a constant WIP_s level at 10 jobs.**

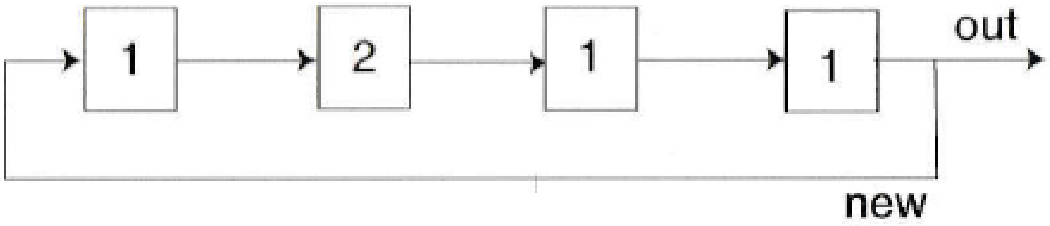


Fig.1.4. A four machine serial flow production factory with constant service times and a constant WIP_s level

(1) **Compute the throughput rate of the factory.**

(2) **Compute the average cycle time and x-factor of the factory.**

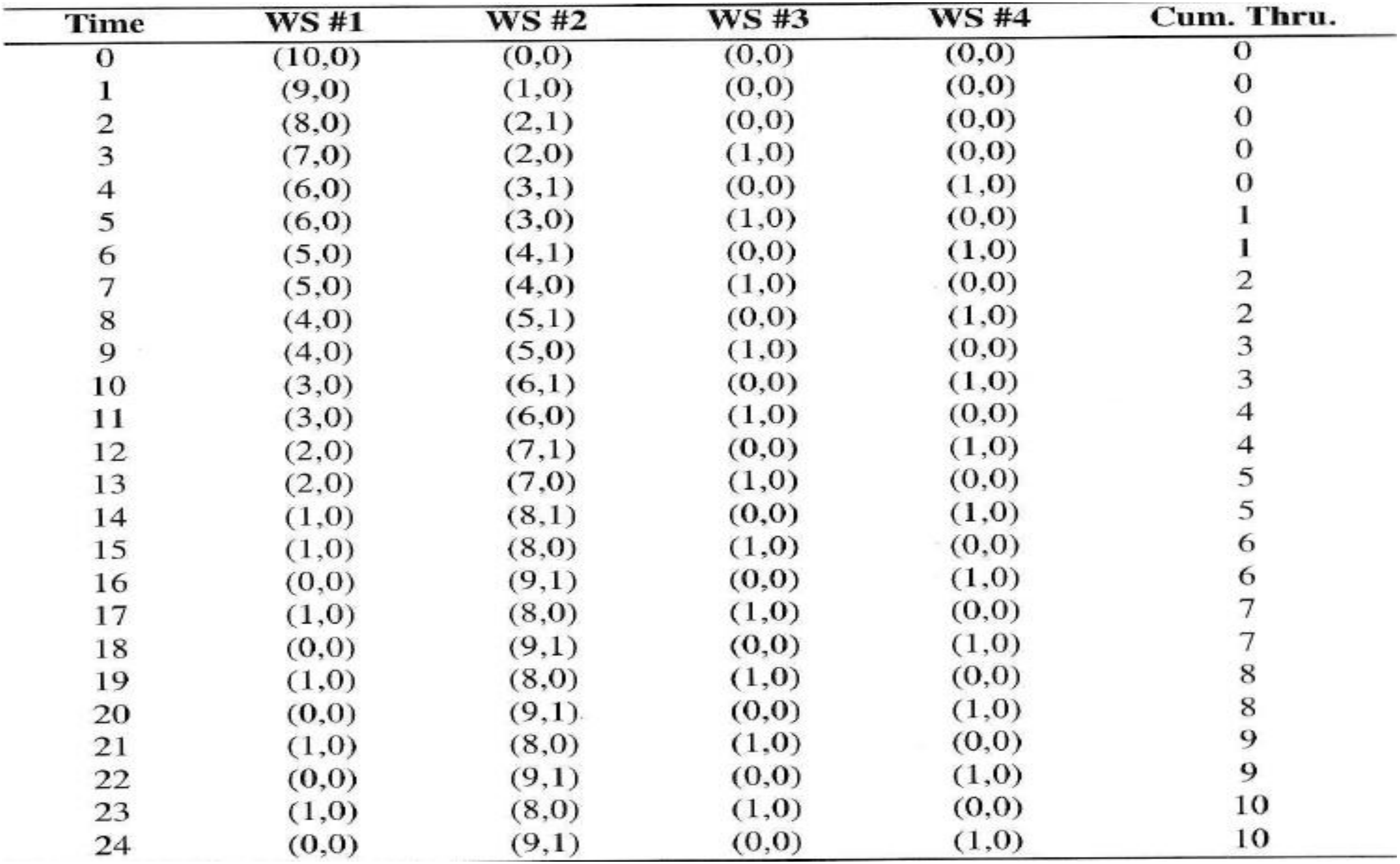
(3) **The average of this industry is currently running at 2.6 as reported in a recent publication by the industry's professional journal. If the x-factor is high, it is difficult to keep customers when the industry on average produces the same product with a considerably shorter lead-time from order placement to receipt.**

Answers:

(1) **Simulate the factory operation. Start with the specified number of 10 jobs in the factory, all placed at Machine 1, and made hourly updates to each job's status. After a short period of time, it accomplishes a throughput rate of *th*=0.5/*hr* jobs (leaving the factory), i.e., it produces one finished job every two hours on the average. This is the maximum throughput rate for this factory because its slowest processing step (at Machine 2) takes two hours per job. Management is quite pleased with the throughput of the factory since it is at its maximum capacity.**

(2) **The cycle time is currently running at 20 hours per job. Management feels like this is high since it takes 5 hours of processing to complete each job. The ratio of the cycle time to the processing time (i.e., x-factor) is 4.**

Table 1.1 Factory simulation with WIP=10, four single-machine workstations, and processing times of (1,2,1,1) for one 24-hour day using a time step of one hour; data pairs under each workstation are the number of jobs at the workstation and the elapsed processing time for the job bing processed.



Time	WS #1	WS #2	WS #3	WS #4	Cum. Thru.
0	(10,0)	(0,0)	(0,0)	(0,0)	0
1	(9,0)	(1,0)	(0,0)	(0,0)	0
2	(8,0)	(2,1)	(0,0)	(0,0)	0
3	(7,0)	(2,0)	(1,0)	(0,0)	0
4	(6,0)	(3,1)	(0,0)	(1,0)	0
5	(6,0)	(3,0)	(1,0)	(0,0)	1
6	(5,0)	(4,1)	(0,0)	(1,0)	1
7	(5,0)	(4,0)	(1,0)	(0,0)	2
8	(4,0)	(5,1)	(0,0)	(1,0)	2
9	(4,0)	(5,0)	(1,0)	(0,0)	3
10	(3,0)	(6,1)	(0,0)	(1,0)	3
11	(3,0)	(6,0)	(1,0)	(0,0)	4
12	(2,0)	(7,1)	(0,0)	(1,0)	4
13	(2,0)	(7,0)	(1,0)	(0,0)	5
14	(1,0)	(8,1)	(0,0)	(1,0)	5
15	(1,0)	(8,0)	(1,0)	(0,0)	6
16	(0,0)	(9,1)	(0,0)	(1,0)	6
17	(1,0)	(8,0)	(1,0)	(0,0)	7
18	(0,0)	(9,1)	(0,0)	(1,0)	7
19	(1,0)	(8,0)	(1,0)	(0,0)	8
20	(0,0)	(9,1)	(0,0)	(1,0)	8
21	(1,0)	(8,0)	(1,0)	(0,0)	9
22	(0,0)	(9,1)	(0,0)	(1,0)	9
23	(1,0)	(8,0)	(1,0)	(0,0)	10
24	(0,0)	(9,1)	(0,0)	(1,0)	10

(3) To decrease x-factor is to decrease the cycle time.

Firstly, management has been considering a large capital outlay to purchase a 25% faster machine (1.5 hours) for processing step two. Then the x-factor decreases to 3.33 and the additional throughput of 0.166 units per hour. However, management has decided that this investment is not worthwhile.

Secondly, management hired a consulting team from the manufacturing engineering department of a local university to perform a short term factory flow analysis study.

Thirdly, the consulting team started to simulate the factory model. The team found a two-hour cyclic pattern. Every cycle of this pattern produced one completed job and the factory returned to the identical state for each machine and associated queue. This set of conditions is referred to as the factory status.

Fourthly, the team used Little's Law to make model analysis.

$$CT = WIP/th.$$

Using a throughput rate of ½ jobs per hour, then cycle time is given by

$$CT = 2 \times WIP\ . \qquad x = \frac{CT}{5} = \frac{WIP}{2.5}\ .$$

So a small *x* will be obtained with a less *WIP*. If *WIP* is 6, *x* is less than 2.6.

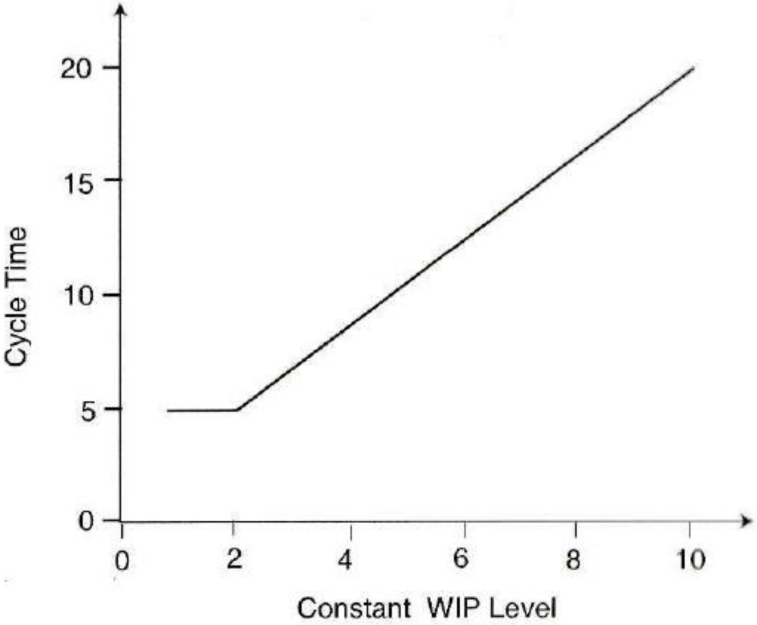


Fig.1.5. Average cycle time as a function of the constant *WIP* level

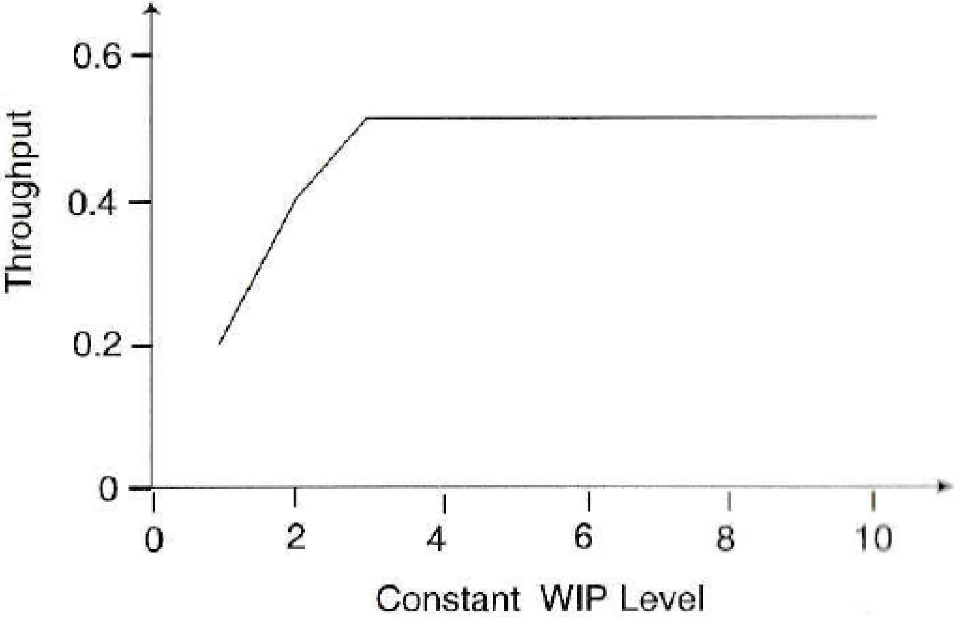
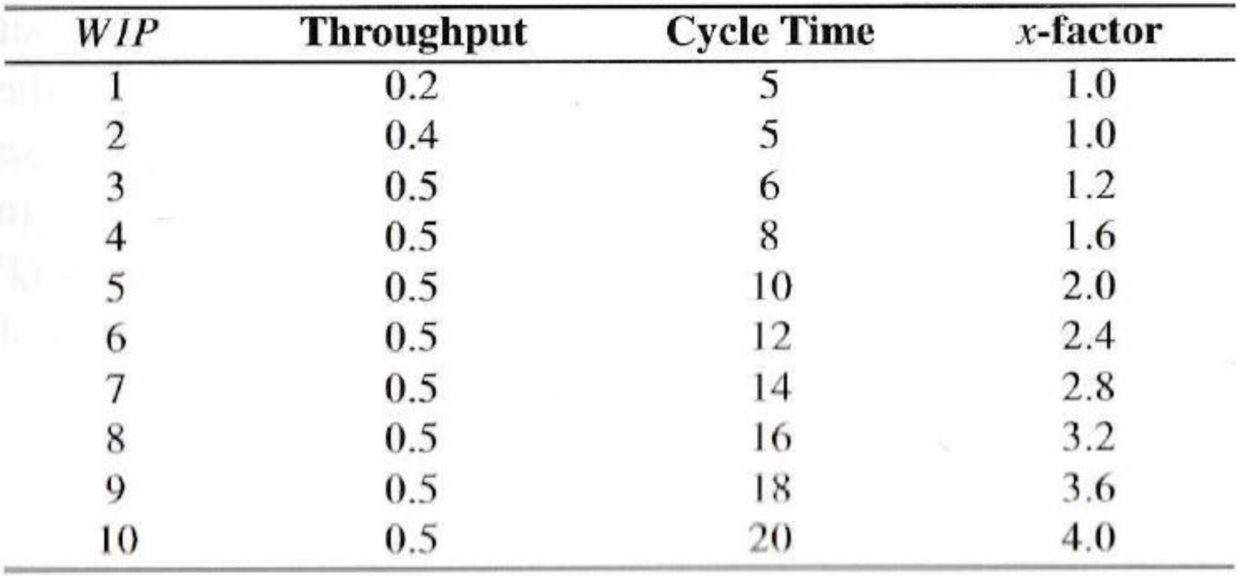


Fig.1.6. Average throughput rate as a function of the constant *WIP* level



WIP	**Throughput**	**Cycle Time**	*x*-**factor**
1	0.2	5	1.0
2	0.4	5	1.0
3	0.5	6	1.2
4	0.5	8	1.6
5	0.5	10	2.0
6	0.5	12	2.4
7	0.5	14	2.8
8	0.5	16	3.2
9	0.5	18	3.6
10	0.5	20	4.0

Table 1.2 Factory performance measures as a function of the WIP level

- **The simple throughput analysis of a serial factory does not necessarily yield accurate results when processing times are random.**
- **Consider the four-step production system again. Now instead of the constant processing time of two hours at workstation 2, let us assume that this time actually varies between 1 hour and 3 hours.**
- **These situations occur at the machine with equal probability for a given job.**

WIP	1 hour	3 hours	Average
3	3/4	1/3	13/24
4	1	1/3	2/3
5	1	1/3	2/3

Table 1.3 Weighted average throughput rate results for the factory with Workstation 2 processing times of 1 and 3 hours, and constant WIP levels of 3,4 and 5.

- **If the proportion of the time that the system is operating in a slow state is 75%, one would expect a more accurate throughput rate estimate to be**

$$3/4(1/3)+1/4(1)=1/2$$

This is the expected throughput rate for the stochastic system if the WIP level is at least the minimum of 4 jobs.

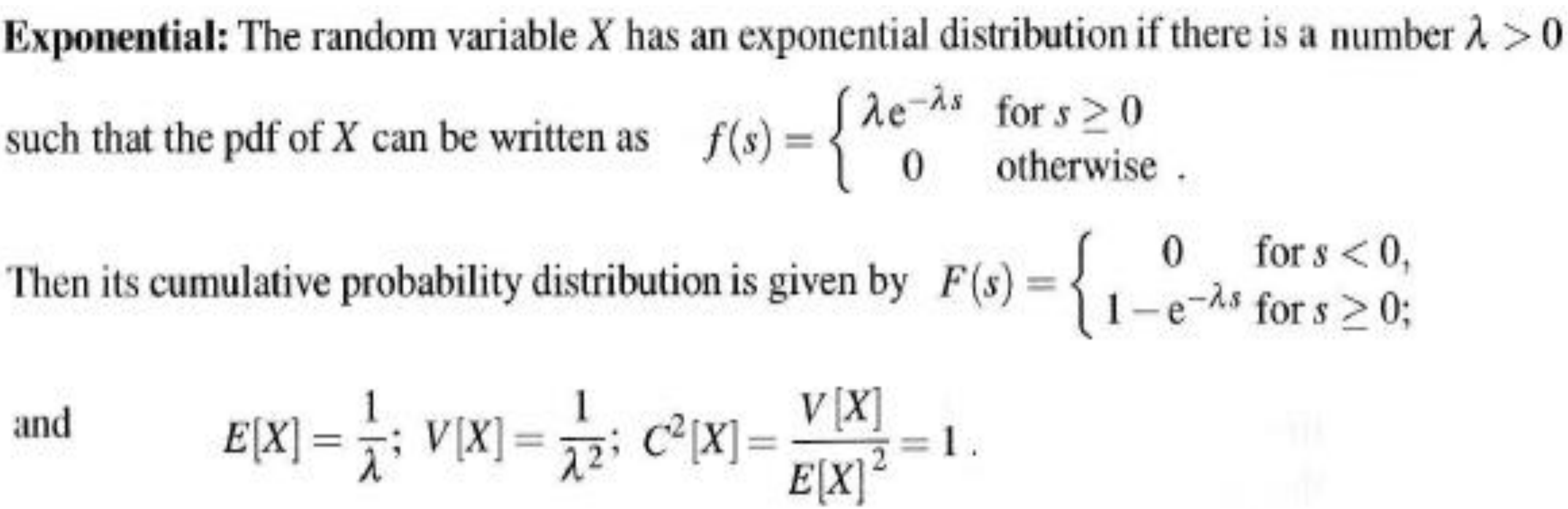
- **Notice the detrimental effect of the variability in the processing time; namely, a necessary increase in WIP and CT to maintain the same throughput rate.**
- **In general, variability in workplace parameters always is detrimental in that it increases average work-in-process and cycle times.**

1.3 Single workstation factory models

- For analysis on the steady-state system measures such as *WIP* and *CT*, it is useful to obtain the probability mass function (*pmf*) of the steady-state number of jobs in the system.

- For notational purposes, define the random variable N as the number of jobs in the system and define p_n as the probability that the number of jobs in the system is n, namely, $p_n=Pr\{N=n\}$.

- It is assumed that the arrival times and processing times of the jobs are subject to exponential distribution in the initial models.



Exponential: The random variable X has an exponential distribution if there is a number $\lambda > 0$ such that the pdf of X can be written as

$$f(s) = \begin{cases} \lambda e^{-\lambda s} & \text{for } s \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Then its cumulative probability distribution is given by

$$F(s) = \begin{cases} 0 & \text{for } s < 0, \\ 1 - e^{-\lambda s} & \text{for } s \geq 0; \end{cases}$$

and

$$E[X] = \frac{1}{\lambda};\ V[X] = \frac{1}{\lambda^2};\ C^2[X] = \frac{V[X]}{E[X]^2} = 1.$$

- Important assumptions
  - Job inter-arrival times are independent of the status of the system.
  - Server will never be idle when there is a job in the system that can be served.
    - Server will be always busy processing jobs when there are jobs available for service.
    - Server will be only idle when there are no jobs available.

1.3.1 First Model

Consider a single server:

- It is with a limited waiting area for n_{max}-1 jobs and one in the server position, i.e., a maximum of n_{max} jobs in the system.
- Jobs arrive to the system one at a time with exponentially distributed inter-arrival times. Denoting the mean arrival rate as λ, the mean inter-arrival time is $1/\lambda$.
- If the system is full, the arriving job is rejected; otherwise, the arriving job is accepted and processed in a first-come-first-serve order.
- The processing time is also assumed to be exponentially distributed, with mean rate μ (the mean service time is $1/\mu$).

It is assumed that a steady-state exists, i.e., the flow into and out of each state are balance. Develop the steady-state distribution of the number of jobs in the system.

Analyze:

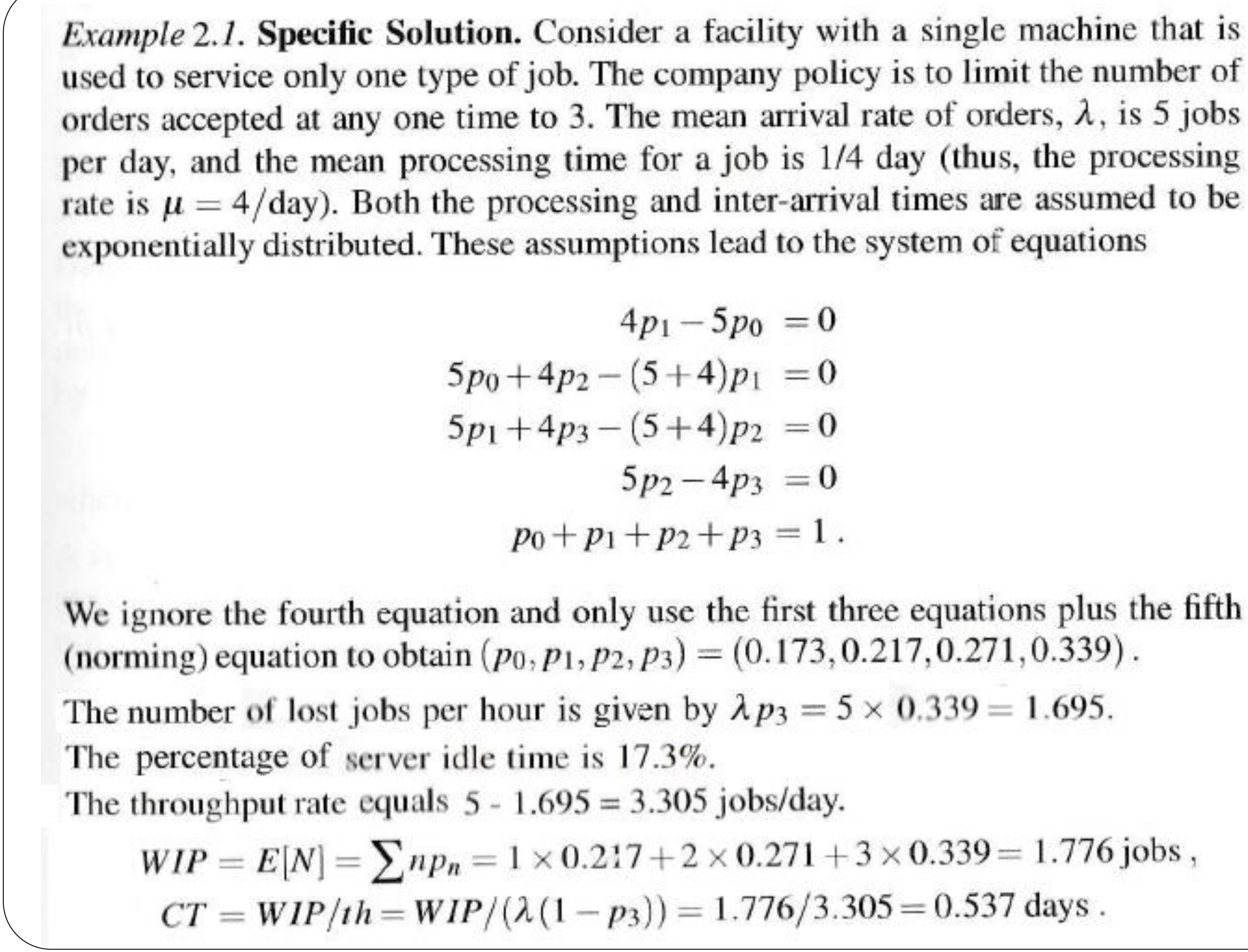
- There are n_{max} +1 possible states, i.e., {0,1,…, n_{max} }, representing the number of jobs in the system.
- Let p_n denote the steady-state probability of n jobs in the system for n=0,…,n_{max}.
- The steady-state flow *into* an intermediate state n ($0<n<n_{max}$) is made up of two components:
  - A new job's arrival to the system that has exactly n-1 jobs
  - The completion of a job's service when the system has exactly n+1 jobs
- The steady-state flow *out* of an intermediate state n ($0<n<n_{max}$) is also made up of two components:
  - The completion of a job's service when the system has exactly n jobs
  - The arrival of a new job to the system when there are exactly n jobs in the system prior to the arrival event.
- The steady-state flow balance equation for an intermediate state n is

inflow

$$\lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n \quad \text{for } n = 1, \cdots, n_{\max}$$

outflow

$$\mu p_1 = \lambda p_0 \qquad \lambda p_{n_{\max}-1} = \mu p_{n_{\max}} . \qquad \sum_{n=0}^{n_{\max}} p_n = 1 .$$



Example 2.1. **Specific Solution.** Consider a facility with a single machine that is used to service only one type of job. The company policy is to limit the number of orders accepted at any one time to 3. The mean arrival rate of orders, λ, is 5 jobs per day, and the mean processing time for a job is 1/4 day (thus, the processing rate is $\mu = 4/\text{day}$). Both the processing and inter-arrival times are assumed to be exponentially distributed. These assumptions lead to the system of equations

$$\begin{aligned} 4p_1 - 5p_0 &= 0 \\ 5p_0 + 4p_2 - (5+4)p_1 &= 0 \\ 5p_1 + 4p_3 - (5+4)p_2 &= 0 \\ 5p_2 - 4p_3 &= 0 \\ p_0 + p_1 + p_2 + p_3 &= 1\,. \end{aligned}$$

We ignore the fourth equation and only use the first three equations plus the fifth (norming) equation to obtain $(p_0, p_1, p_2, p_3) = (0.173, 0.217, 0.271, 0.339)$.

The number of lost jobs per hour is given by $\lambda p_3 = 5 \times 0.339 = 1.695$.

The percentage of server idle time is 17.3%.

The throughput rate equals 5 - 1.695 = 3.305 jobs/day.

$$WIP = E[N] = \sum n p_n = 1 \times 0.217 + 2 \times 0.271 + 3 \times 0.339 = 1.776 \text{ jobs}\,,$$

$$CT = WIP/th = WIP/(\lambda(1 - p_3)) = 1.776/3.305 = 0.537 \text{ days}\,.$$

Example 2.2. General Solution.

$$\begin{aligned}
\mu p_1 - \lambda p_0 &= 0 \\
\lambda p_0 + \mu p_2 - (\lambda+\mu) p_1 &= 0 \\
\lambda p_1 + \mu p_3 - (\lambda+\mu) p_2 &= 0 \\
\lambda p_2 - \mu p_3 &= 0 \\
p_0 + p_1 + p_2 + p_3 &= 1 \,.
\end{aligned}$$

➡

$$\begin{aligned}
\mu p_1 &= \lambda p_0 \\
p_1 &= \frac{\lambda}{\mu} p_0 \,.
\end{aligned}$$

➡

$$\begin{aligned}
\lambda p_0 + \mu p_2 &= (\lambda+\mu) p_1 \\
\mu p_2 &= (\lambda+\mu) p_1 - \lambda p_0 \\
p_2 &= (\lambda+\mu) \frac{\lambda}{\mu^2} p_0 - \frac{\lambda}{\mu} p_0 \\
p_2 &= \left(\frac{\lambda}{\mu}\right)^2 p_0 \,.
\end{aligned}$$

$$\begin{aligned}
\lambda p_1 + \mu p_3 &= (\lambda+\mu) p_2 \\
p_3 &= (\lambda+\mu) \frac{\lambda^2}{\mu^3} p_0 - \left(\frac{\lambda}{\mu}\right)^2 p_0 \\
p_3 &= \left(\frac{\lambda}{\mu}\right)^3 p_0 \,.
\end{aligned}$$

$$\begin{aligned}
1 &= p_0 + p_1 + p_2 + p_3 \\
&= \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right] p_0 = 1
\end{aligned}$$

$$p_0 = \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^{3}\right]^{-1} .$$

From here we can develop the measures of $WIP = p_1 + 2p_2 + 3p_3$, $th = \lambda(p_0 + p_1 + p_2)$, and $CT = WIP/th$. □

Whenever the system is finite, there is the possibility that the system will be full and arriving jobs will be lost, hence, the actual rate of jobs that enter the system, λ_e may not be the same as the arrival rate, λ.

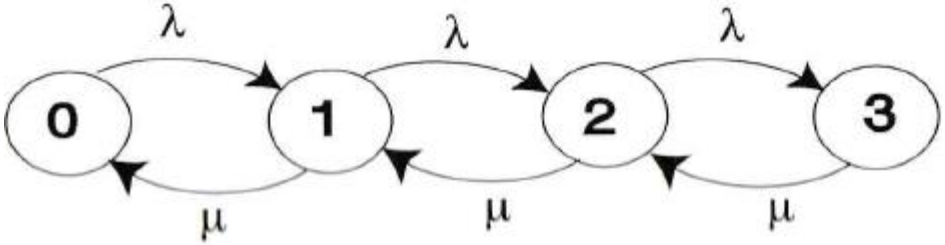
Definition 2.1. The *effective arrival rate* for a system is the rate at which jobs enter the system. For a workstation with constant arrival rate, λ, and with a maximum number of jobs at the workstation limited to nmax, the effective arrival rate is given by

$$\lambda_e=\lambda(1-p_{nmax})$$

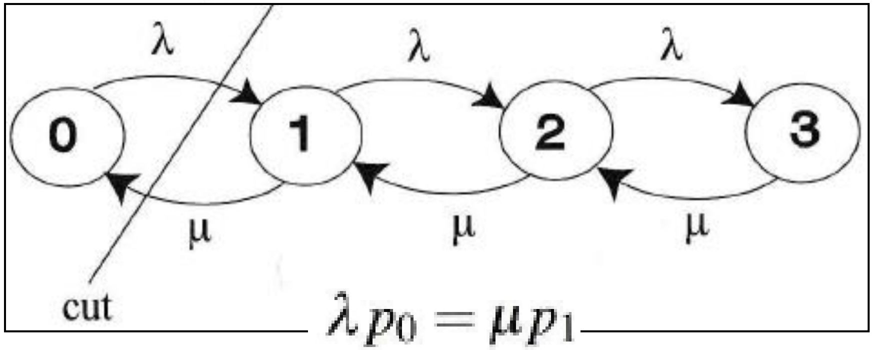
where p_{nmax} is the probability that the workstation is full.

A system at steady-state will have its system throughput rate equal to the effective arrival rate; that is, $th=\lambda_e$, and the use of Little's Law must always use λ_e and not λ for the throughput.

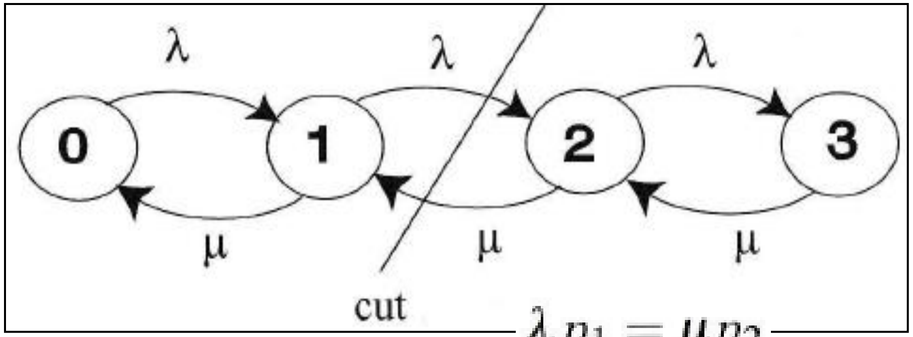
1.3.2 Diagram Method for Developing the Balance Equations



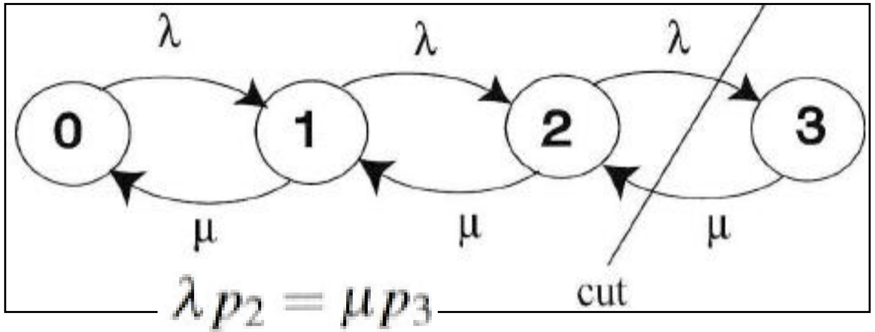
Partition the nodes into two subsets of nodes, then establish values for the appropriate steady-state probabilities to balance the flow between the two subsets.



$\lambda p_0 = \mu p_1$



$\lambda p_1 = \mu p_2$

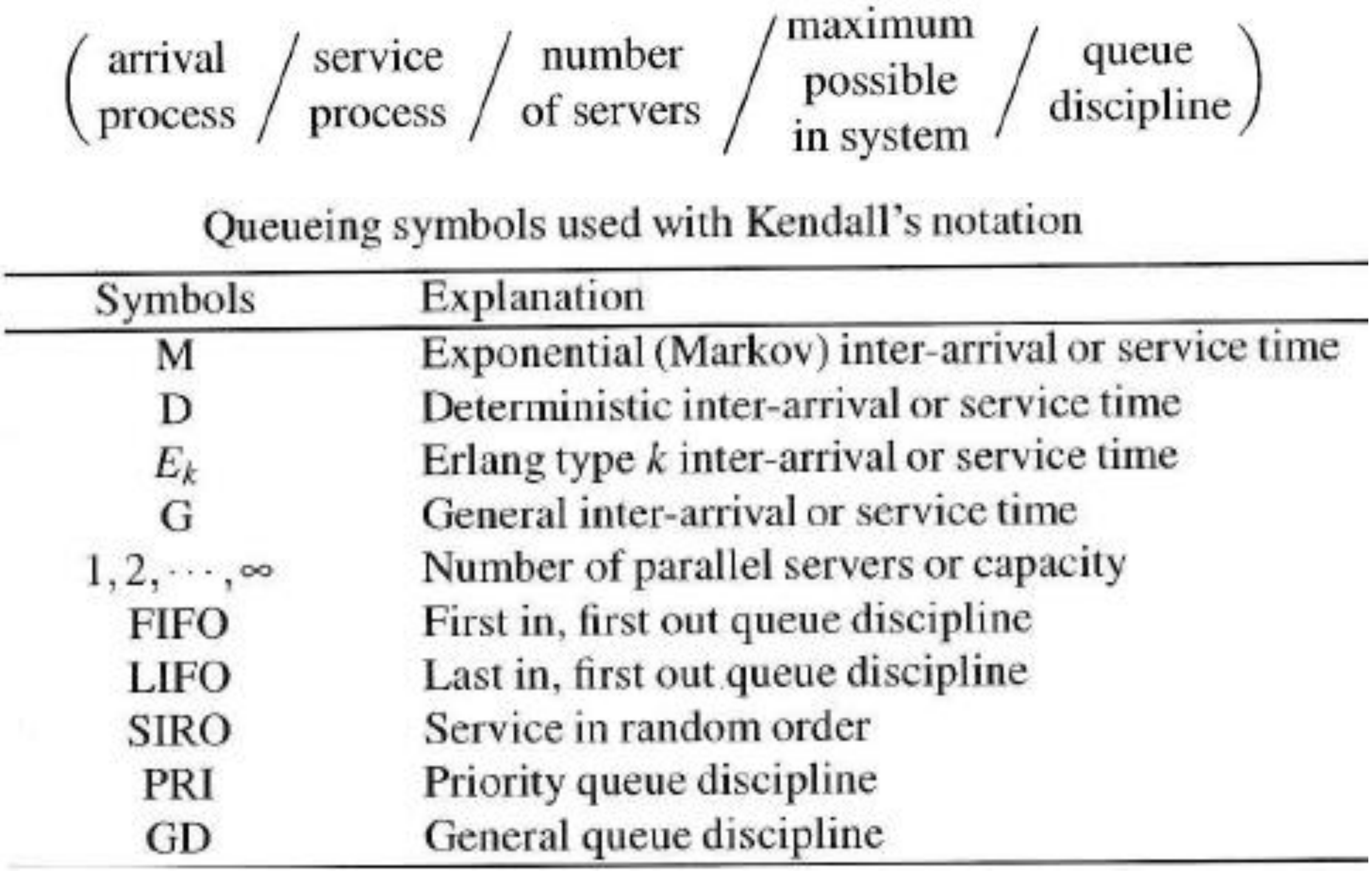


$\lambda p_2 = \mu p_3$

$$\lambda p_0 = \mu p_1$$
$$\lambda p_1 = \mu p_2$$
$$\lambda p_2 = \mu p_3$$
$$\sum_{n=0}^{3} p_n = 1.$$

1.3.3 Model Shorthand Notation

- The general form of Kendall’s notation is



$$\left(\begin{array}{c}\text{arrival}\\ \text{process}\end{array} \Big/ \begin{array}{c}\text{service}\\ \text{process}\end{array} \Big/ \begin{array}{c}\text{number}\\ \text{of servers}\end{array} \Big/ \begin{array}{c}\text{maximum}\\ \text{possible}\\ \text{in system}\end{array} \Big/ \begin{array}{c}\text{queue}\\ \text{discipline}\end{array}\right)$$

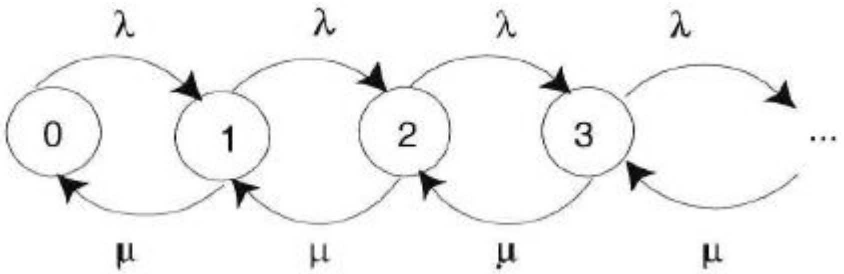
Queueing symbols used with Kendall’s notation

Symbols	Explanation
M	Exponential (Markov) inter-arrival or service time
D	Deterministic inter-arrival or service time
E_k	Erlang type *k* inter-arrival or service time
G	General inter-arrival or service time
$1, 2, \cdots, \infty$	Number of parallel servers or capacity
FIFO	First in, first out queue discipline
LIFO	Last in, first out queue discipline
SIRO	Service in random order
PRI	Priority queue discipline
GD	General queue discipline

1.3.4. An Infinite Capacity Model (*M/M/*1)

- The effective arrival rate (those jobs getting into the system) will necessarily be less than the system's service capacity. For a given M/M/1/3 system,
  - With $\lambda=\mu$, $p_0=\ldots=p_3=1/4$, $\lambda e= \lambda(1-p_3)=(3/4)\ \lambda< \mu$
  - With $\lambda=2\mu$, $p_0=(1/2)p_1=(1/4)\ p_2=(1/8)p_3$, $\lambda e= \lambda(1-p_3)=(7/15)\ \lambda< \mu$
  - With $\lambda=3\mu$, $p_0=(1/3)p_1=(1/9)\ p_2=(1/27)p_3$, $\lambda_e= \lambda(1-p_3)=(13/40)\ \lambda< \mu$
- Note that as the ratio of λ/μ becomes larger, the effective arrival rate approaches the inverse of this ratio but never reaches it.
- The finite capacity systems have a built-in mechanism to adjust the arrival rate to a level (λ_e) that can be handled by the system service capacity.
- If a system that has no realistic limit on the number of jobs allowed is considered, no steady-state exists.
- The analyses of the unlimited queuing models result in conditions that establish the existence of the steady-state behavior for these model.

- The formulation of the unlimited-jobs system



- Using the cut partitions method for obtaining system of equations needed in defining the steady-state probabilities

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ \lambda p_1 &= \mu p_2 \\ \lambda p_2 &= \mu p_3 \\ &\vdots \\ \lambda p_n &= \mu p_{n+1} \\ &\vdots \\ \sum_{n=0}^{\infty} p_n &= 1\,. \end{aligned}$$

⇒

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu} p_0 \\ p_2 &= \frac{\lambda}{\mu} p_1 \\ p_3 &= \frac{\lambda}{\mu} p_2 \\ &\vdots \\ p_n &= \frac{\lambda}{\mu} p_{n-1} \\ &\vdots \end{aligned}$$

⇒

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad \text{for } n = 0, 1, \cdots .$$

$$p_0 + \left(\frac{\lambda}{\mu}\right) p_0 + \left(\frac{\lambda}{\mu}\right)^2 p_0 + \cdots + \left(\frac{\lambda}{\mu}\right)^n p_0 + \cdots = 1\,,$$

$$p_0 = \frac{1}{\left(1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \cdots + \left(\frac{\lambda}{\mu}\right)^n + \cdots\right)}\,.$$

The denominator is a geometric series that has a finite value if $\lambda/\mu < 1$.

$$p_0 = 1 - \frac{\lambda}{\mu}$$

and the general solution to the steady-state probabilities is (given that $\lambda/\mu < 1$)

$$p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \quad \text{for } n = 0, 1, \cdots .$$

The throughput rate per unit time for this system is λ. The utilization factor u for the server is obtained from

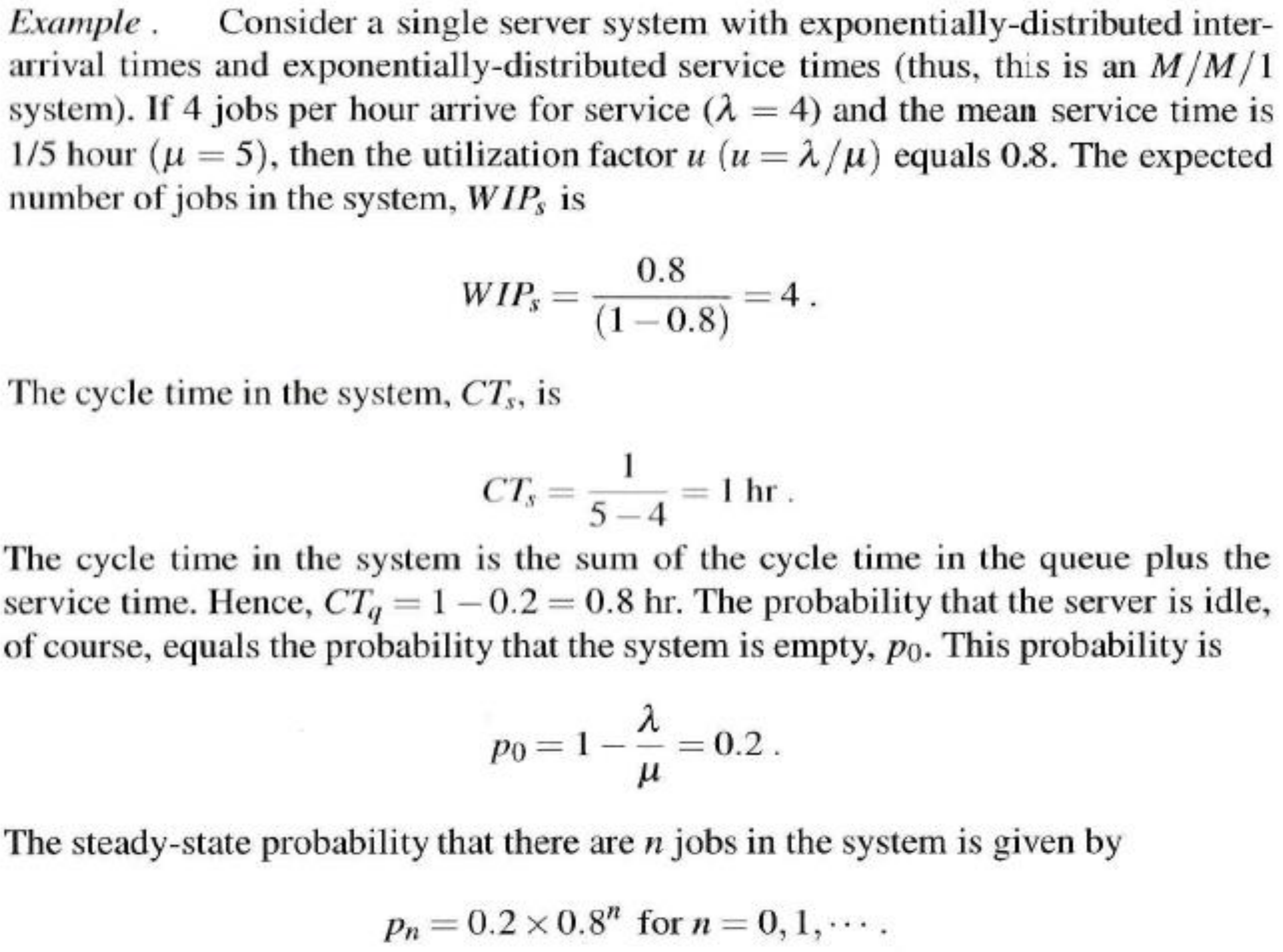
$$u = 1 - p_0 = 1 - \left(1 - \frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu} .$$

The expected number of jobs in the system in steady-state is obtained by using the derivative of the geometric series as follows:

$$
\begin{aligned}
WIP_s = E[N] = \sum_{n=0}^{\infty} np_n &= \sum_{n=0}^{\infty} n\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \\
&= \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)\sum_{n=1}^{\infty} n\left(\frac{\lambda}{\mu}\right)^{n-1} \\
&= \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)\left(\frac{1}{1-\frac{\lambda}{\mu}}\right)^2 \\
&= \frac{\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)^2} = \frac{\frac{\lambda}{\mu}}{\left(1-\frac{\lambda}{\mu}\right)} = \frac{u}{1-u}
\end{aligned}
$$

where N is a random variable denoting the number of jobs in the system. Using Little's Law, the expected time in system (the cycle time) CT_s is given by

$$CT_s = \frac{WIP_s}{\lambda} = \frac{1}{\lambda}\frac{\frac{\lambda}{\mu}}{(1-\frac{\lambda}{\mu})} = \frac{1}{\mu-\lambda}\,.$$



Example . Consider a single server system with exponentially-distributed inter-arrival times and exponentially-distributed service times (thus, this is an $M/M/1$ system). If 4 jobs per hour arrive for service ($\lambda = 4$) and the mean service time is 1/5 hour ($\mu = 5$), then the utilization factor u ($u = \lambda/\mu$) equals 0.8. The expected number of jobs in the system, WIP_s is

$$WIP_s = \frac{0.8}{(1-0.8)} = 4\,.$$

The cycle time in the system, CT_s, is

$$CT_s = \frac{1}{5-4} = 1 \text{ hr}\,.$$

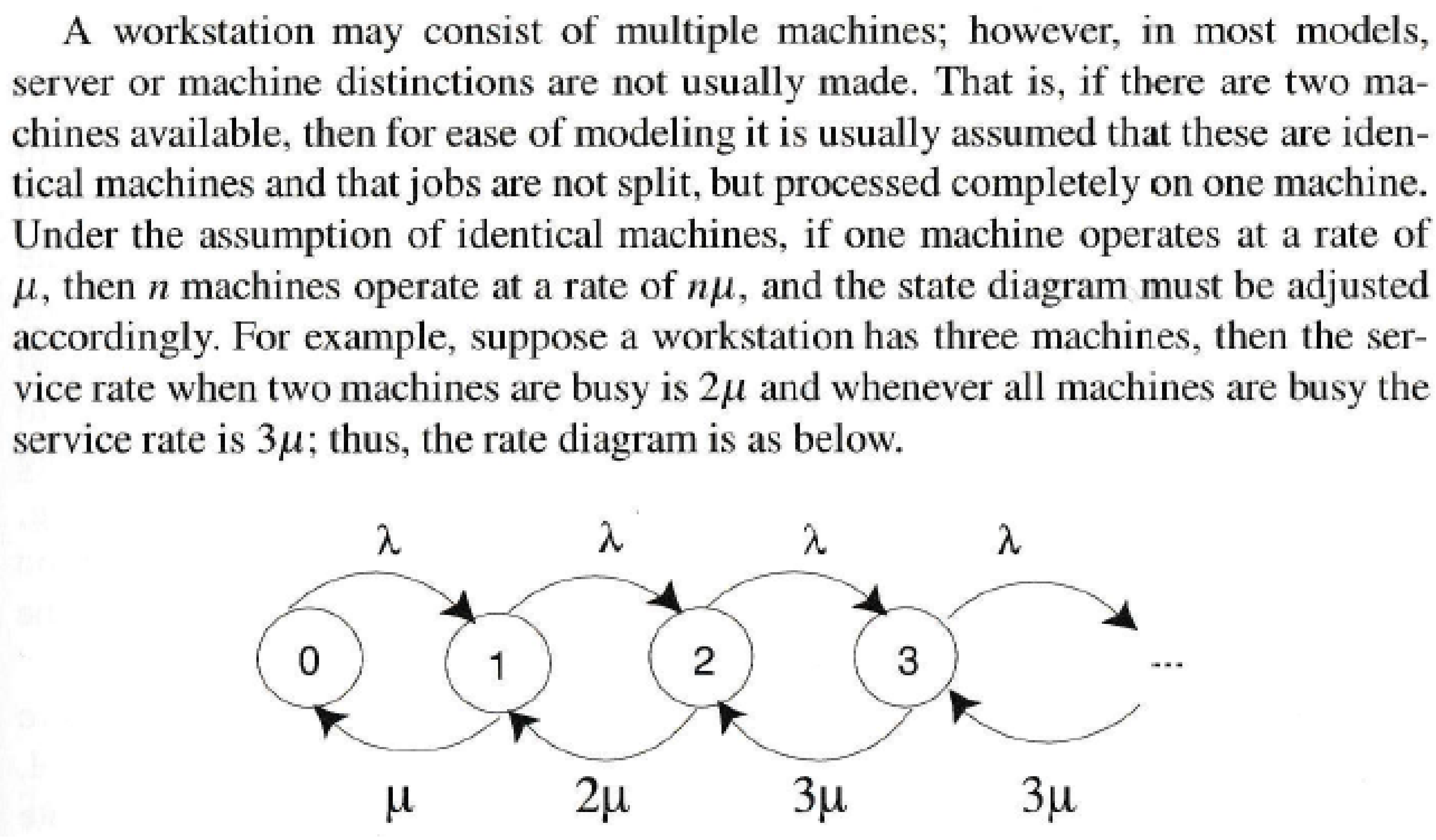
The cycle time in the system is the sum of the cycle time in the queue plus the service time. Hence, $CT_q = 1 - 0.2 = 0.8$ hr. The probability that the server is idle, of course, equals the probability that the system is empty, p_0. This probability is

$$p_0 = 1 - \frac{\lambda}{\mu} = 0.2\,.$$

The steady-state probability that there are n jobs in the system is given by

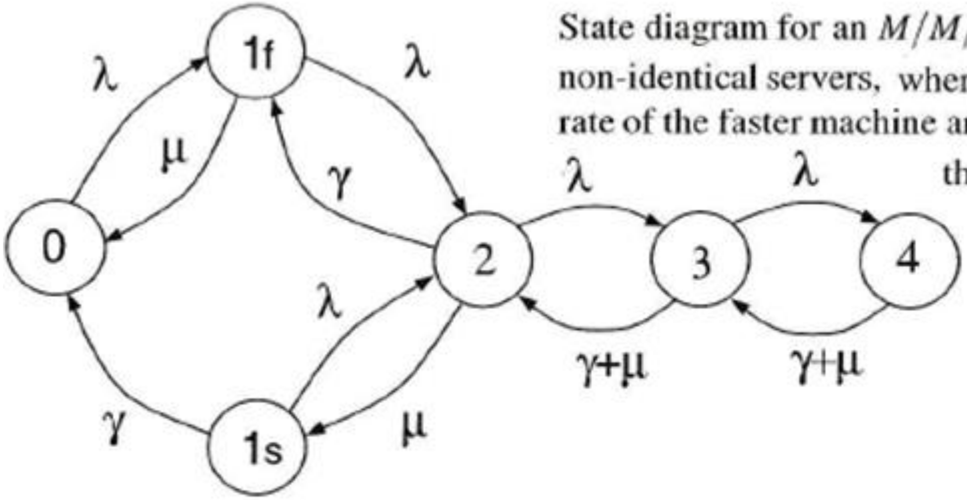
$$p_n = 0.2 \times 0.8^n \text{ for } n = 0, 1, \cdots\,.$$

- **Multiple Server Systems with Identical Service Rates**



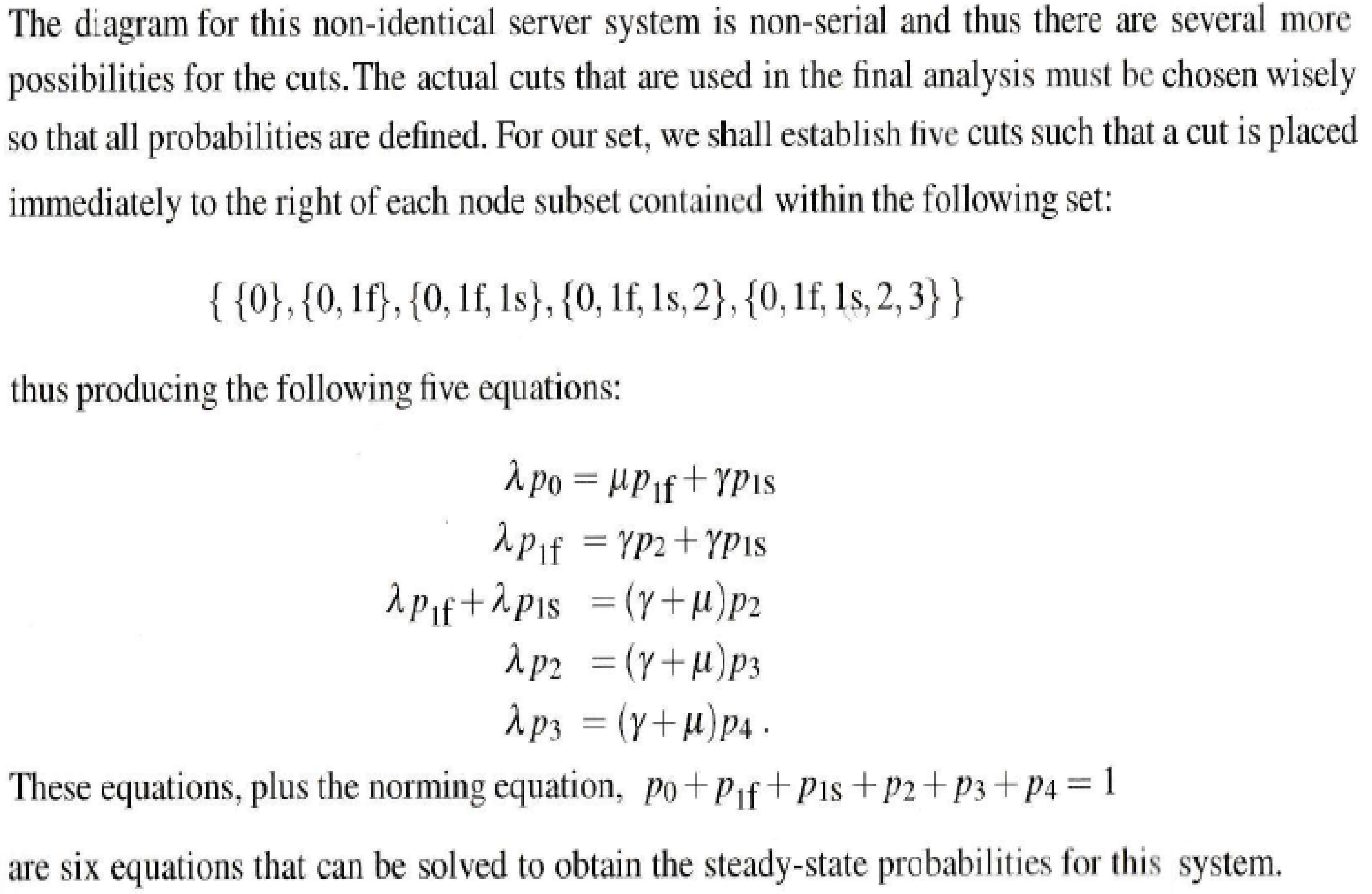
A workstation may consist of multiple machines; however, in most models, server or machine distinctions are not usually made. That is, if there are two machines available, then for ease of modeling it is usually assumed that these are identical machines and that jobs are not split, but processed completely on one machine. Under the assumption of identical machines, if one machine operates at a rate of μ, then n machines operate at a rate of $n\mu$, and the state diagram must be adjusted accordingly. For example, suppose a workstation has three machines, then the service rate when two machines are busy is 2μ and whenever all machines are busy the service rate is 3μ; thus, the rate diagram is as below.

1.3.5 Multiple Server Systems with Non-Identical Service Rates



State diagram for an $M/M/2/4$ system with non-identical servers, where μ denotes the rate of the faster machine and γ is the rate of the slower machine

As before, $n_{\max}$ is the maximum number of jobs allowed in the system (here $n_{\max} = 4$) so that there will be a total of $n_{\max} + 2$ possible states for this model. In the identical server model, there were $n_{\max} + 1$ possible states. The extra state arises because we must know which machine is busy when there is only one job at the workstation in order to know the service rate associated with the job in process. When there are two or more jobs in the system, both machines are busy and no distinction about the state needs to be made. Denoting the state (i.e., the number of jobs at the workstation) by n, one possible state space is the set $\{0, 1\text{f}, 1\text{s}, 2, 3, 4\}$, where $n = 1\text{f}$ indicates that one job is in the system and that job is being processed on the fast machine and $n = 1\text{s}$ indicates that one job is in the system and is being processed on the slow machine.



The diagram for this non-identical server system is non-serial and thus there are several more possibilities for the cuts. The actual cuts that are used in the final analysis must be chosen wisely so that all probabilities are defined. For our set, we shall establish five cuts such that a cut is placed immediately to the right of each node subset contained within the following set:

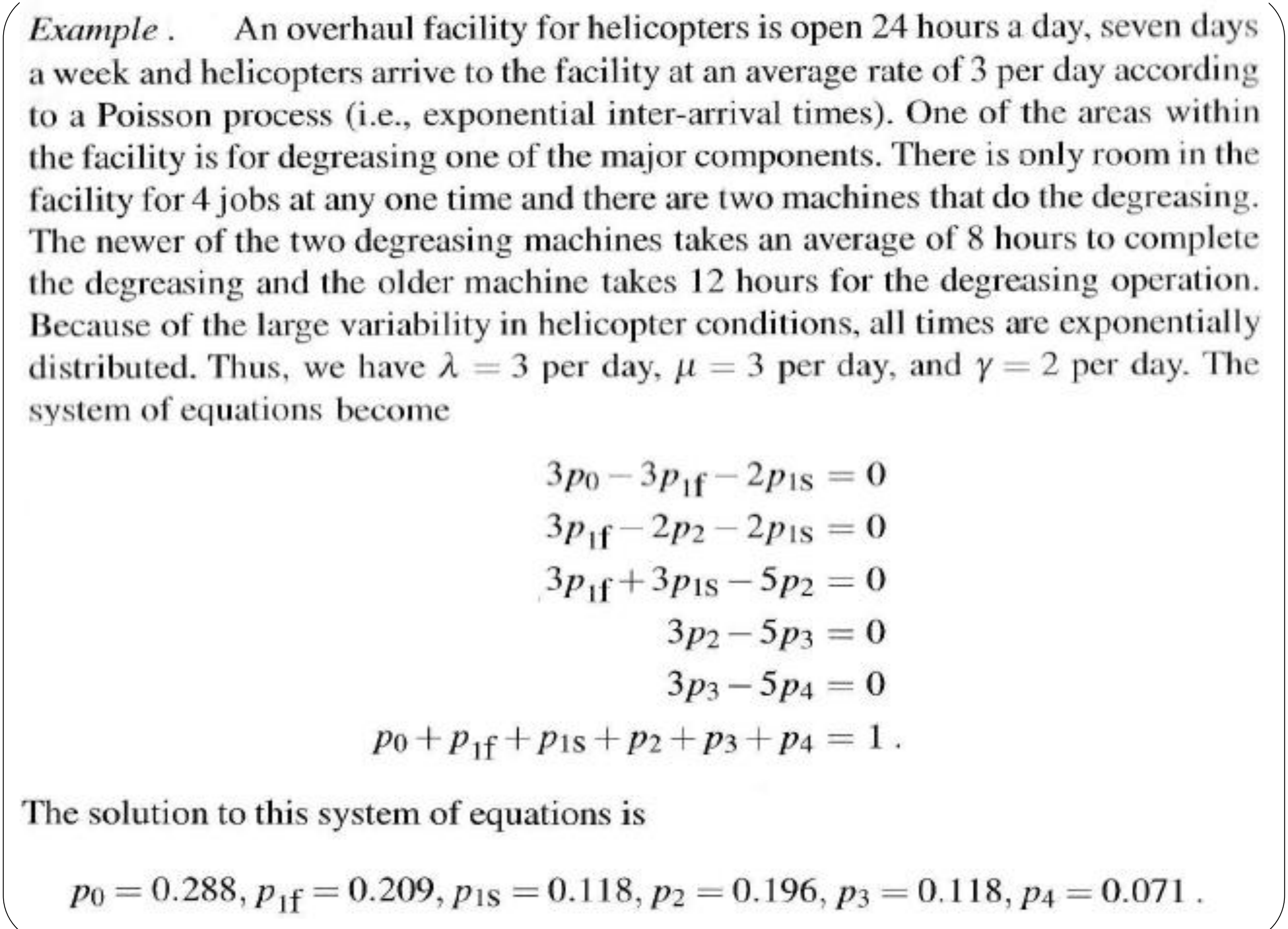
$$\{\,\{0\},\{0,1\mathrm{f}\},\{0,1\mathrm{f},1\mathrm{s}\},\{0,1\mathrm{f},1\mathrm{s},2\},\{0,1\mathrm{f},1\mathrm{s},2,3\}\,\}$$

thus producing the following five equations:

$$\begin{aligned}
\lambda p_0 &= \mu p_{1\mathrm{f}} + \gamma p_{1\mathrm{s}} \\
\lambda p_{1\mathrm{f}} &= \gamma p_2 + \gamma p_{1\mathrm{s}} \\
\lambda p_{1\mathrm{f}} + \lambda p_{1\mathrm{s}} &= (\gamma+\mu) p_2 \\
\lambda p_2 &= (\gamma+\mu) p_3 \\
\lambda p_3 &= (\gamma+\mu) p_4 \,.
\end{aligned}$$

These equations, plus the norming equation, $p_0 + p_{1\mathrm{f}} + p_{1\mathrm{s}} + p_2 + p_3 + p_4 = 1$

are six equations that can be solved to obtain the steady-state probabilities for this system.

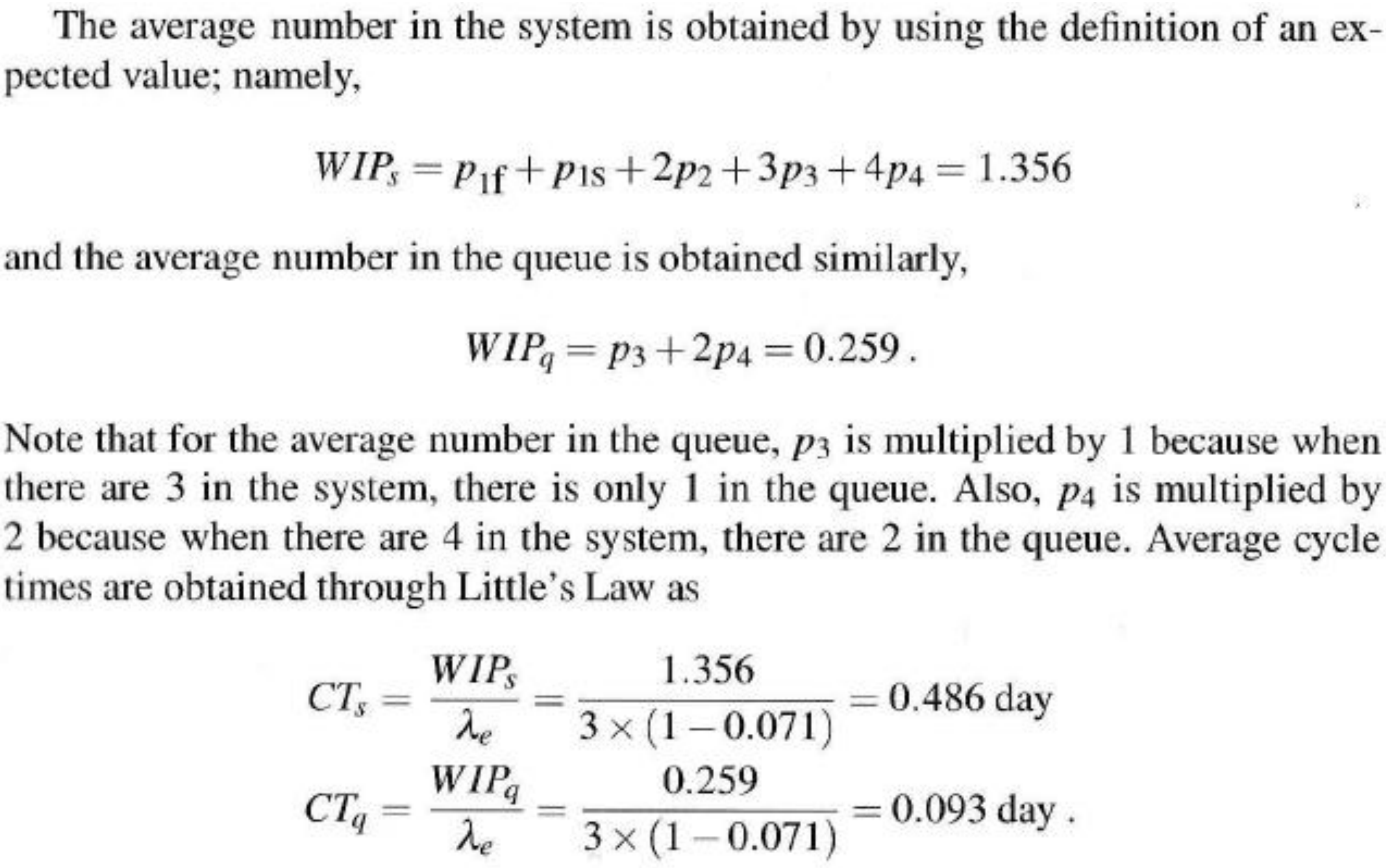


Example . An overhaul facility for helicopters is open 24 hours a day, seven days a week and helicopters arrive to the facility at an average rate of 3 per day according to a Poisson process (i.e., exponential inter-arrival times). One of the areas within the facility is for degreasing one of the major components. There is only room in the facility for 4 jobs at any one time and there are two machines that do the degreasing. The newer of the two degreasing machines takes an average of 8 hours to complete the degreasing and the older machine takes 12 hours for the degreasing operation. Because of the large variability in helicopter conditions, all times are exponentially distributed. Thus, we have $\lambda = 3$ per day, $\mu = 3$ per day, and $\gamma = 2$ per day. The system of equations become

$$\begin{aligned}
3p_0 - 3p_{1f} - 2p_{1s} &= 0\\
3p_{1f} - 2p_2 - 2p_{1s} &= 0\\
3p_{1f} + 3p_{1s} - 5p_2 &= 0\\
3p_2 - 5p_3 &= 0\\
3p_3 - 5p_4 &= 0\\
p_0 + p_{1f} + p_{1s} + p_2 + p_3 + p_4 &= 1\,.
\end{aligned}$$

The solution to this system of equations is

$$p_0 = 0.288, p_{1f} = 0.209, p_{1s} = 0.118, p_2 = 0.196, p_3 = 0.118, p_4 = 0.071\,.$$



The average number in the system is obtained by using the definition of an expected value; namely,

$$WIP_s = p_{1\mathrm{f}} + p_{1\mathrm{s}} + 2p_2 + 3p_3 + 4p_4 = 1.356$$

and the average number in the queue is obtained similarly,

$$WIP_q = p_3 + 2p_4 = 0.259\,.$$

Note that for the average number in the queue, p_3 is multiplied by 1 because when there are 3 in the system, there is only 1 in the queue. Also, p_4 is multiplied by 2 because when there are 4 in the system, there are 2 in the queue. Average cycle times are obtained through Little's Law as

$$CT_s = \frac{WIP_s}{\lambda_e} = \frac{1.356}{3 \times (1 - 0.071)} = 0.486 \text{ day}$$

$$CT_q = \frac{WIP_q}{\lambda_e} = \frac{0.259}{3 \times (1 - 0.071)} = 0.093 \text{ day}\,.$$

A couple of other measures that are sometimes desired by management are the number of busy processors (i.e., degreasers) and their utilization. The expected number of busy servers, $E[BS]$, is 1.097, and is obtained as

$$E[BS] = 1p_{1\mathrm{f}} + 1p_{1\mathrm{s}} + 2p_2 + 2P_3 + 2p_4 = 1.097 .$$

The system utilization factor u is the expected number of busy servers divided by the number of machines available

$$u = \frac{E[BS]}{2} = 0.5485 = 54.85\% .$$

Our final calculation is to obtain the average time needed for degreasing. Because of the preference given to using the faster machine, we would expect the average time to be closer to 8 hours than to 12 hours. To get an exact value, we take advantage of the fact that the time in the system equals the time in the queue plus service time

$$E[T] = CT_s - CT_q = 0.486 - 0.093 = 0.393 \text{ days} = 9.4 \text{ hr} .$$

1.3.6 Using Exponentials to Approximate General Times

- To model more general systems, one effective method is to approximate the general times by combinations of exponentials.
- Erlang-*k* distribution, the sum of *k* independent and identical exponential distributions, provides an excellent distribution to use for the expanded state modeling approach.
  - The non-negative random variable *X* has an Erlang distribution if there is a positive integer *k* and a positive number *β* such that the *pdf* of *X* can be written as

$$f(s) = \frac{k(ks)^{k-1}\mathrm{e}^{-(k/\beta)s}}{\beta^k (k-1)!} \quad \text{for } s \geq 0 \, . \qquad E[X] = \beta; \; V[X] = \frac{\beta^2}{k}; \; C^2[X] = \frac{1}{k} \, .$$

  - The Erlang-*k* distribution can be modeled as a serial k-node system, with each node referring to identical exponentials.
  - Erlang-*k* has a squared coefficient of variation given by C^2=1/*k*, it also allows modeling of processes that have less variation than the exponential distribution.

(1) Erlang Processing Times

Consider a single server:

- The number of jobs allows into the system is limited to three, i.e., $n_{max} = 3$.
- Jobs arrive to the system one at a time with exponentially distributed inter-arrival times. Denoting the mean arrival rate as λ, the mean inter-arrival time is $1/\lambda$.
- The processing time is described by and Erlang-2 distribution with mean rate μ and thus mean service time is $1/\mu$.
- The model is denoted by $M/E_2/1/3$.

This Erlang-2 distribution will be modeled using two exponential nodes (phases), where each node has a mean rate of 2μ.

- Each individual node is exponential.
- The service process will have two nodes representing the two phases of the Erlang-2 distribution

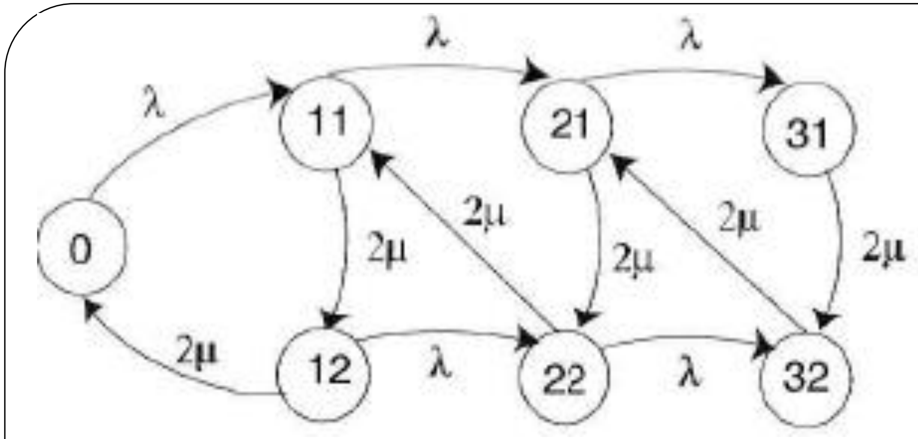


Diagram for an $M/E_2/1/3$ model where the state (n,i) indicates that there are n jobs in the system with the i^{th} service phase busy

$$2n_{\max} + 1 \text{ states}$$

- To obtain the steady-state probabilities for this system, six cuts are placed so that the following node sets are isolated on one side of the cut

$$\{\,\{0\},\{0,(1,2)\},\{0,(1,1)\},\{0,(1,1),(1,2)\},\{(3,1),(3,2)\},\{(3,2)\}\,\}$$

$$\begin{aligned}
\lambda p_0 - 2\mu p_{12} &= 0\\
\lambda p_0 + \lambda p_{12} - 2\mu p_{11} &= 0\\
(\lambda + 2\mu)p_{11} - 2\mu p_{12} - 2\mu p_{22} &= 0\\
\lambda p_{11} + \lambda p_{12} - 2\mu p_{22} &= 0\\
\lambda p_{21} + \lambda p_{22} - 2\mu p_{32} &= 0\\
\lambda p_{22} + 2\mu p_{31} - 2\mu p_{32} &= 0\\
p_0 + p_{11} + p_{12} + p_{21} + p_{22} + p_{31} + p_{32} &= 1\,.
\end{aligned}$$

The performance measures of *WIP*, *CT* and Throughput are computed from

$$WIP_s = \sum_{n=1}^{3} n(p_{n1} + p_{n2})$$

$$th = \lambda_e = \lambda(1 - p_{31} - p_{32})$$

$$CT_s = WIP_s/\lambda_e\,.$$

(2) Erlang Inter-Arrival Times

- Jobs arrive to the system one at a time with Erlang-2 distribution *λ*.
- The processing time is an exponential distribution with mean rate *μ*.
- The model is denoted by $E_2/M/1/3$.

$$\{(1,0), (2,0), (1,1), (2,1), (1,2), (2,2), (1,3), (2,3)\}.$$

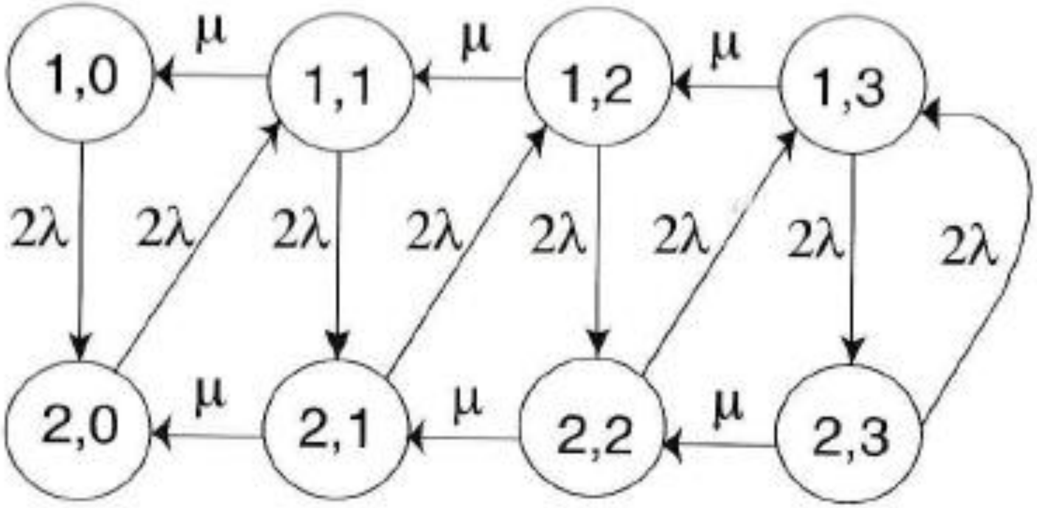


Diagram for an $E_2/M/1/3$ model where the state (i,n) indicates that the arrival process is in phase i and there are n total jobs in the system

$$
\begin{aligned}
2\lambda p_{10} &= \mu p_{11} \\
2\lambda p_{20} &= 2\lambda p_{10} + \mu p_{21} \\
(2\lambda + \mu)p_{11} &= 2\lambda p_{20} + \mu p_{12} \\
(2\lambda + \mu)p_{21} &= 2\lambda p_{11} + \mu p_{22} \\
(2\lambda + \mu)p_{12} &= 2\lambda p_{21} + \mu p_{13} \\
(2\lambda + \mu)p_{22} &= 2\lambda p_{12} + \mu p_{23} \\
(2\lambda + \mu)p_{13} &= 2\lambda p_{22} + 2\lambda p_{23} \\
(2\lambda + \mu)p_{23} &= 2\lambda p_{13}
\end{aligned}
$$

$$p_{10} + p_{20} + p_{11} + p_{21} + p_{12} + p_{22} + p_{13} + p_{23} = 1 .$$

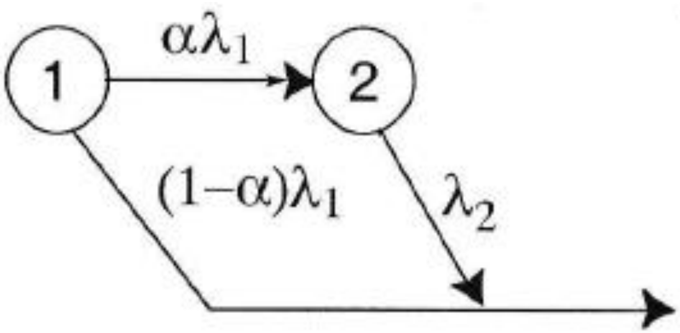
Example. Let $\lambda = 5$ jobs/hr and $\mu = 5$ jobs/hr, and the solution to the $E_2/M/1/3$ system of equations is

$$\begin{aligned} p_{10} &= 0.0687\,, & p_{20} &= 0.1358\,, \\ p_{11} &= 0.1374\,, & p_{21} &= 0.1342\,, \\ p_{12} &= 0.1406\,, & p_{22} &= 0.1278\,, \\ p_{13} &= 0.1534\,, & p_{23} &= 0.1022\,. \end{aligned}$$

Some of the system performance measures are

$$\begin{aligned} WIP_s &= 0(p_{10}+p_{20})+1(p_{11}+p_{21})+2(p_{12}+p_{22})+3(p_{13}+p_{23}) = 1.5751 \\ u &= p_{11}+p_{21}+p_{12}+p_{22}+p_{13}+p_{23} = 1-(p_{10}+p_{20}) = 79.55\% \\ th &= \lambda_e = \lambda - 2\lambda p_{23} = \mu \times u = 3.978 \text{ jobs/hr} \\ CT_s &= WIP_s/th = 0.3960 \text{ hr}\,. \end{aligned}$$

(3) Phased Inter-arrival and Processing Times

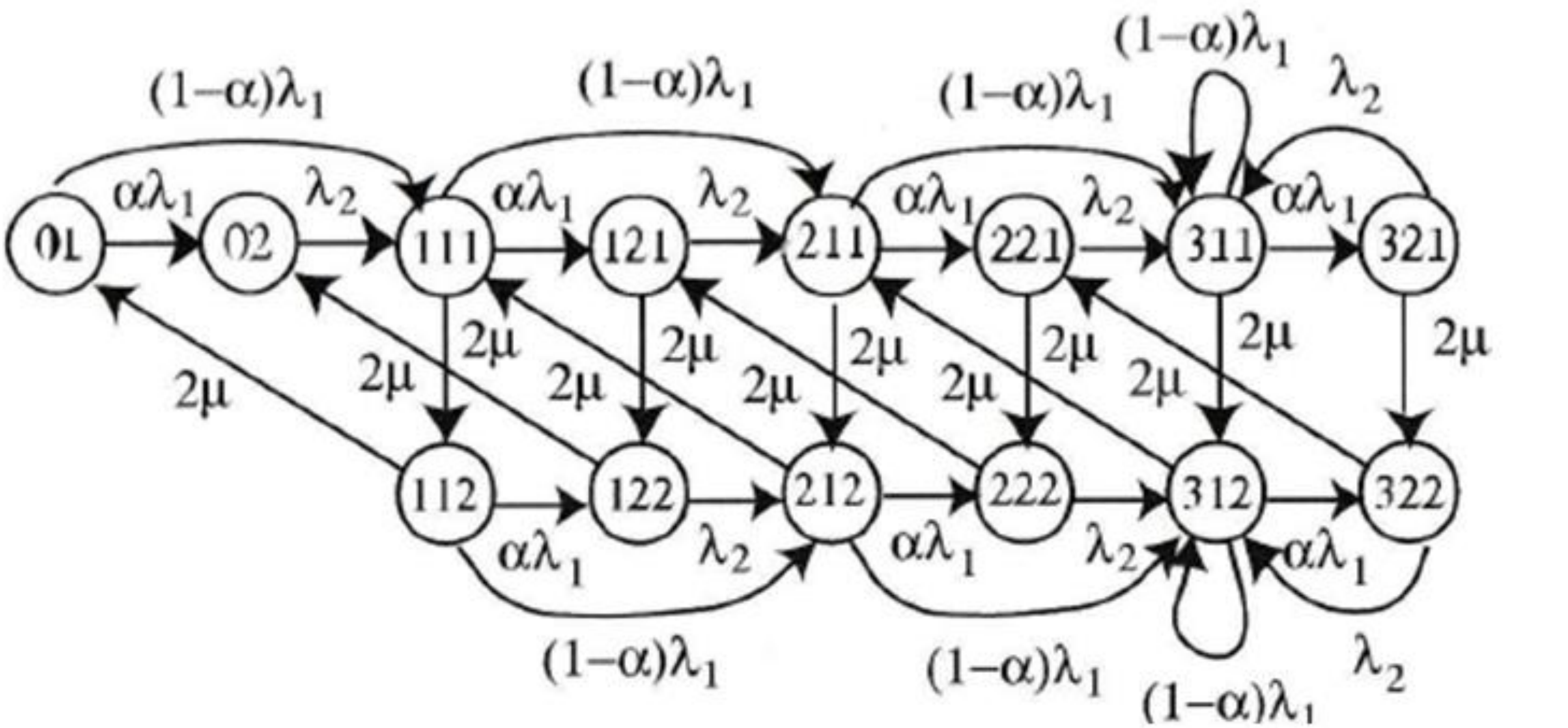


A generalized Erlang with two phases, where the first phase always occurs and has a mean rate λ_1 and the second phase occurs with probability α and has a mean rate λ_2

- Two-phase *GE* (*GE₂*)
  - The first generalization is to allow for non-identical phases and second is to give a probability that the process is complete at the end of each phase.
  - The purpose is to develop modeling skills that have more flexibility in the range of inter-arrival and service time distribution.
  - GE_2 results in a squared coefficient of variation C_2 in the range [0.5, ∞)

$$\lambda_1 = \frac{2}{E[X]}, \quad \lambda_2 = \frac{1}{E[X]C^2[X]}, \quad \alpha = \frac{1}{2C^2[X]} \quad \text{for } C^2[X] > 1 ;$$

$$\lambda_1 = \frac{1}{E[X]C^2[X]}, \quad \lambda_2 = \frac{2}{E[X]}, \quad \alpha = 2(1-C^2[X]) \quad \text{for } \frac{1}{2} \le C^2[X] \le 1 .$$



State diagram for an $GE_2/E_2/1/3$ model, where a (n, i, j) indicates that there are n jobs in the system with one job in arrival phase i and one job is service phase j

1.3.7 Single Server Model Approximations

(1) General Service Distributions

- The Pollaczek and Khintchine, or "*P-K*", formula for *WIP* in an *M*/*G*/1 queuing system is given by

$$WIP_s = E[N] = \frac{\lambda}{\mu} + \frac{(\frac{\lambda}{\mu})^2 + \lambda^2 \sigma_s{}^2}{2(1-\frac{\lambda}{\mu})}$$

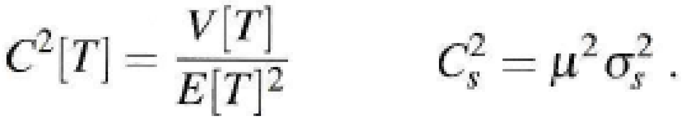
where *N* is the number of jobs in the system, λ is the mean arrival rate, and the service distribution has mean and variance give by 1/μ and $\sigma_s{}^2$, respectively.

$WIP_s - WIP_q = \lambda_e / \mu$ $\quad$ $\lambda_e = \lambda$ for $M/G/1$ $\quad$ $WIP_q = E[N_q] = \dfrac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2 \sigma_s^2}{2\left(1-\frac{\lambda}{\mu}\right)}$.

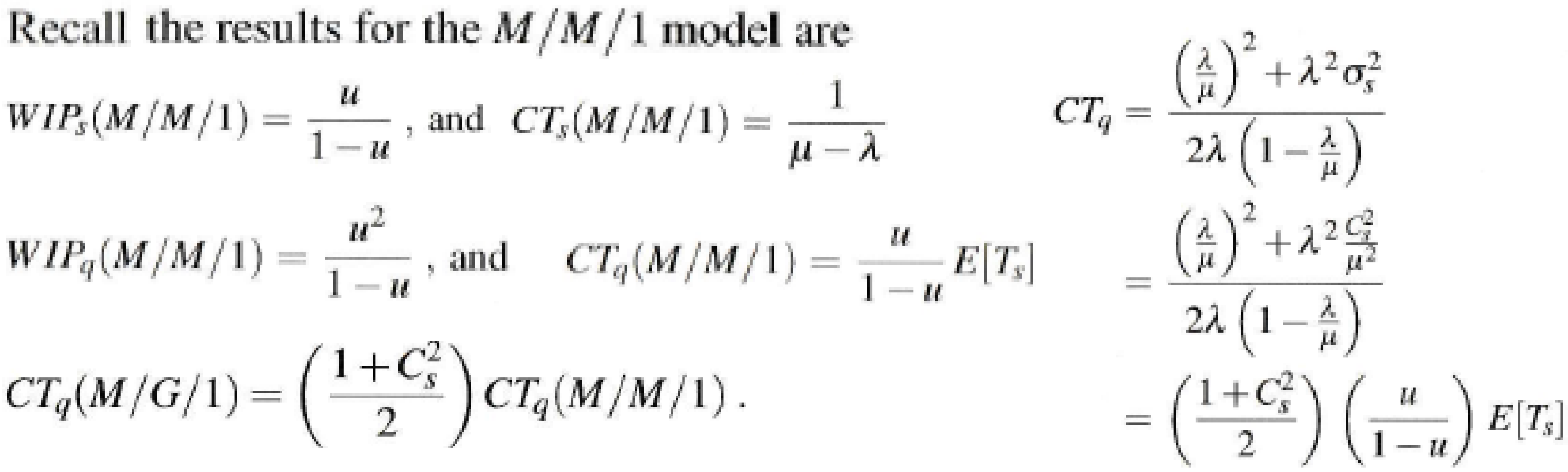
- The P-K formula for the queue cycle time in an M/G/1 system is given by

$$CT_q = \boldsymbol{E}\left[T_q\right] = \frac{WIP_q}{\lambda} = \frac{(\frac{\lambda}{\mu})^2+\lambda^2{\sigma_s}^2}{2\lambda(1-\frac{\lambda}{\mu})}$$

Where T_q is a random variable denoting the time a job spends in the queue, λ is the mean arrival rate, and the service distribution has mean and variance given by 1/μ and ${\sigma_s}^2$, respectively.



$$C^2[T] = \frac{V[T]}{E[T]^2} \qquad C_s^2 = \mu^2\sigma_s^2 .$$



Recall the results for the $M/M/1$ model are

$$WIP_s(M/M/1) = \frac{u}{1-u}\text{, and } CT_s(M/M/1) = \frac{1}{\mu-\lambda}$$

$$WIP_q(M/M/1) = \frac{u^2}{1-u}\text{, and } CT_q(M/M/1) = \frac{u}{1-u}E[T_s]$$

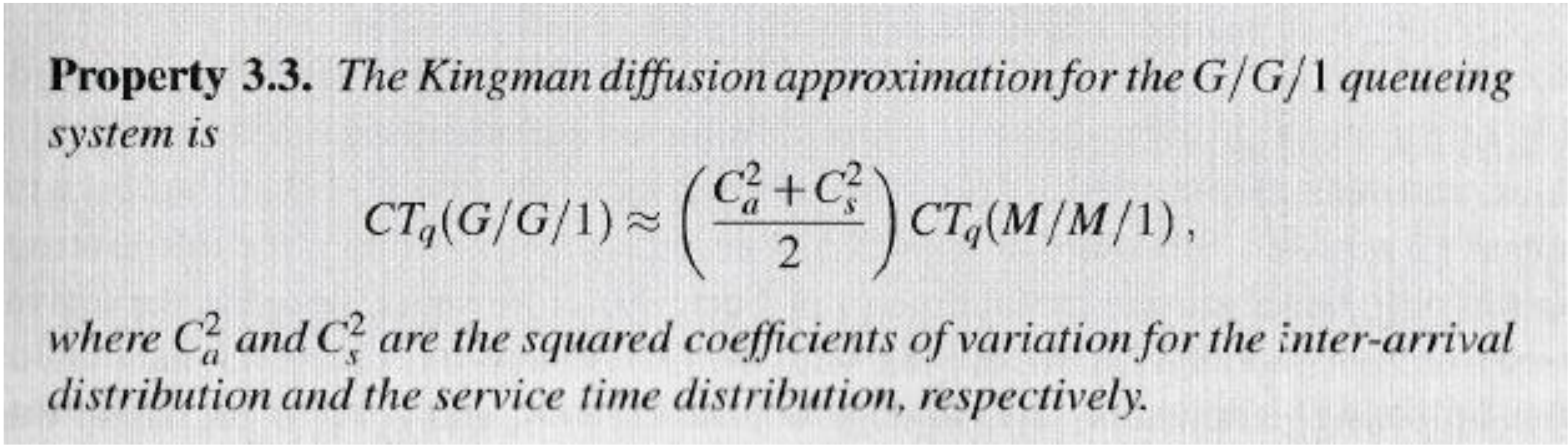
$$CT_q(M/G/1) = \left(\frac{1+C_s^2}{2}\right)CT_q(M/M/1) .$$

$$CT_q = \frac{\left(\frac{\lambda}{\mu}\right)^2+\lambda^2\sigma_s^2}{2\lambda\left(1-\frac{\lambda}{\mu}\right)}$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^2+\lambda^2\frac{C_s^2}{\mu^2}}{2\lambda\left(1-\frac{\lambda}{\mu}\right)}$$

$$= \left(\frac{1+C_s^2}{2}\right)\left(\frac{u}{1-u}\right)E[T_s]$$

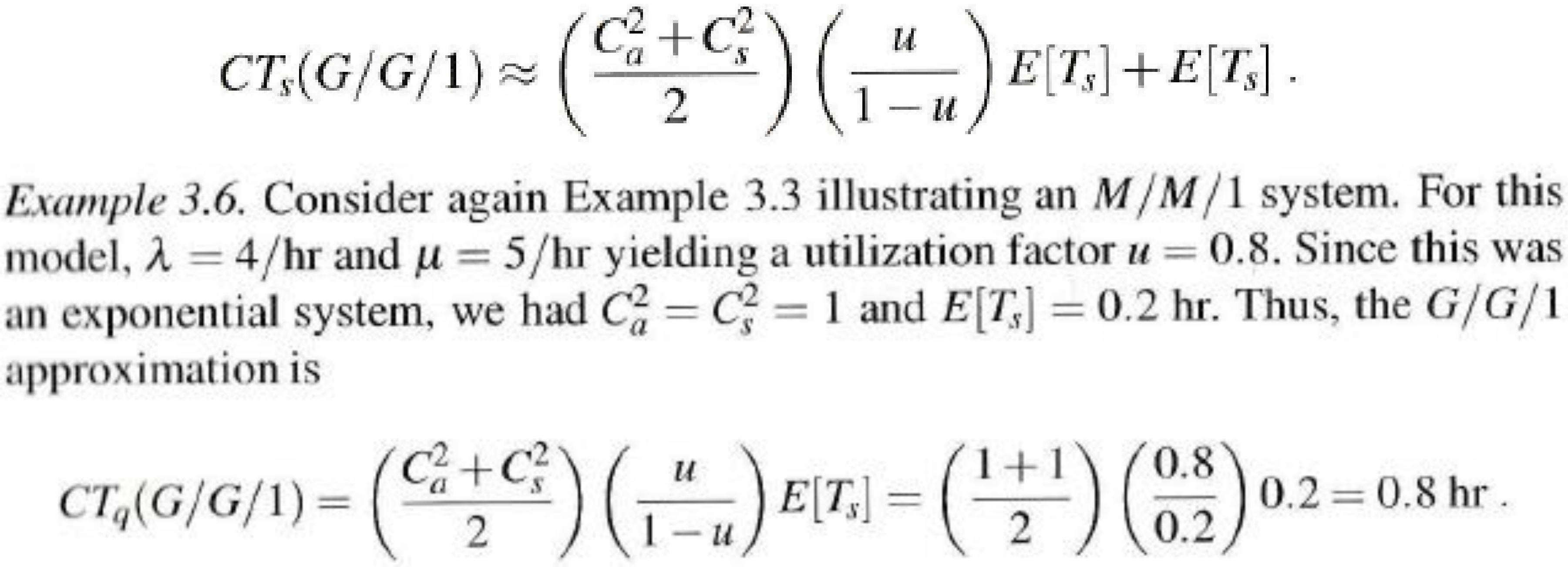
(2) Approximations for G/G/1 System



Property 3.3. *The Kingman diffusion approximation for the $G/G/1$ queueing system is*

$$CT_q(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2}\right) CT_q(M/M/1)\,,$$

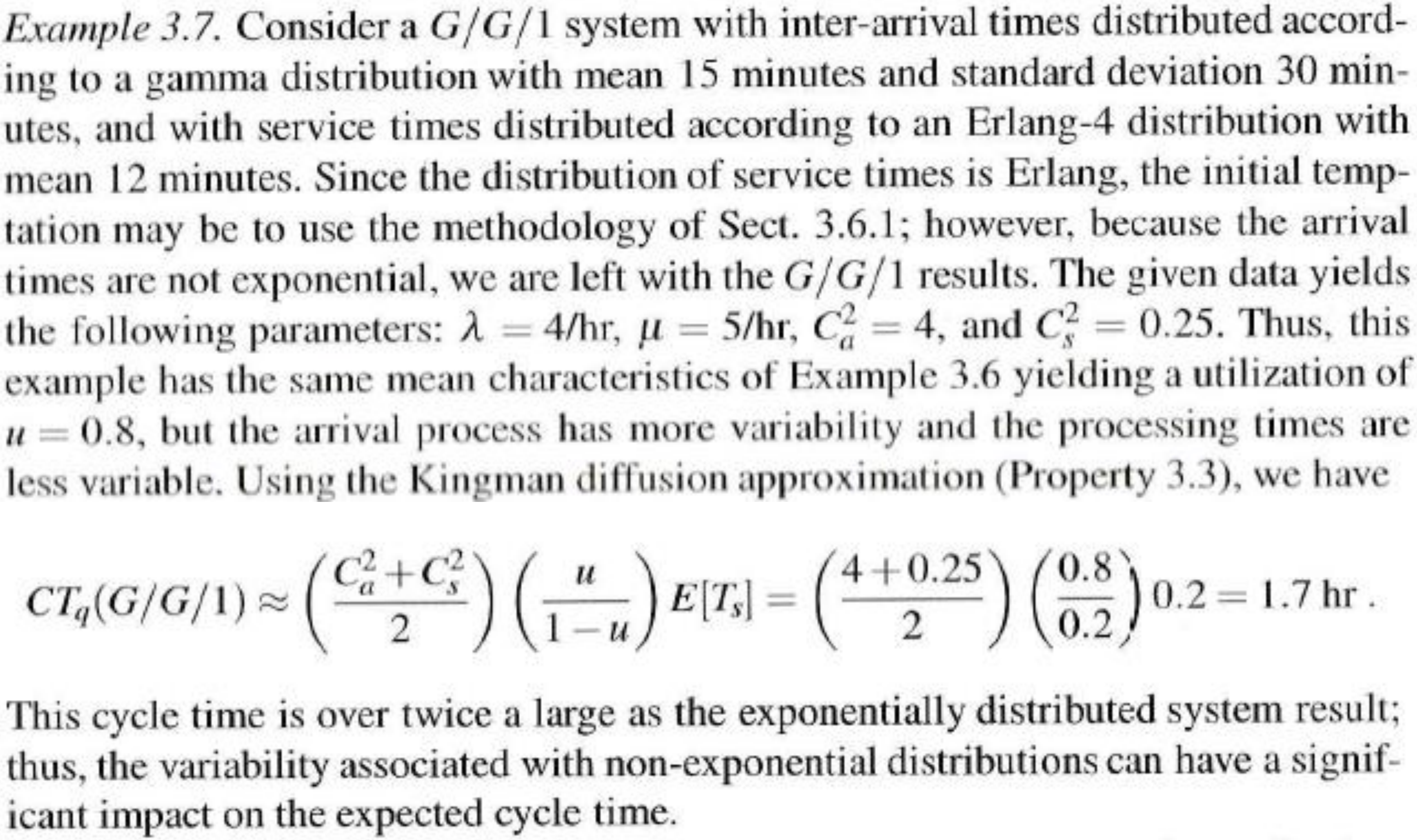
where C_a^2 and C_s^2 are the squared coefficients of variation for the inter-arrival distribution and the service time distribution, respectively.



$$CT_s(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right) E[T_s] + E[T_s]\,.$$

Example 3.6. Consider again Example 3.3 illustrating an $M/M/1$ system. For this model, $\lambda = 4/\text{hr}$ and $\mu = 5/\text{hr}$ yielding a utilization factor $u = 0.8$. Since this was an exponential system, we had $C_a^2 = C_s^2 = 1$ and $E[T_s] = 0.2$ hr. Thus, the $G/G/1$ approximation is

$$CT_q(G/G/1) = \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right) E[T_s] = \left(\frac{1+1}{2}\right)\left(\frac{0.8}{0.2}\right) 0.2 = 0.8 \text{ hr}\,.$$

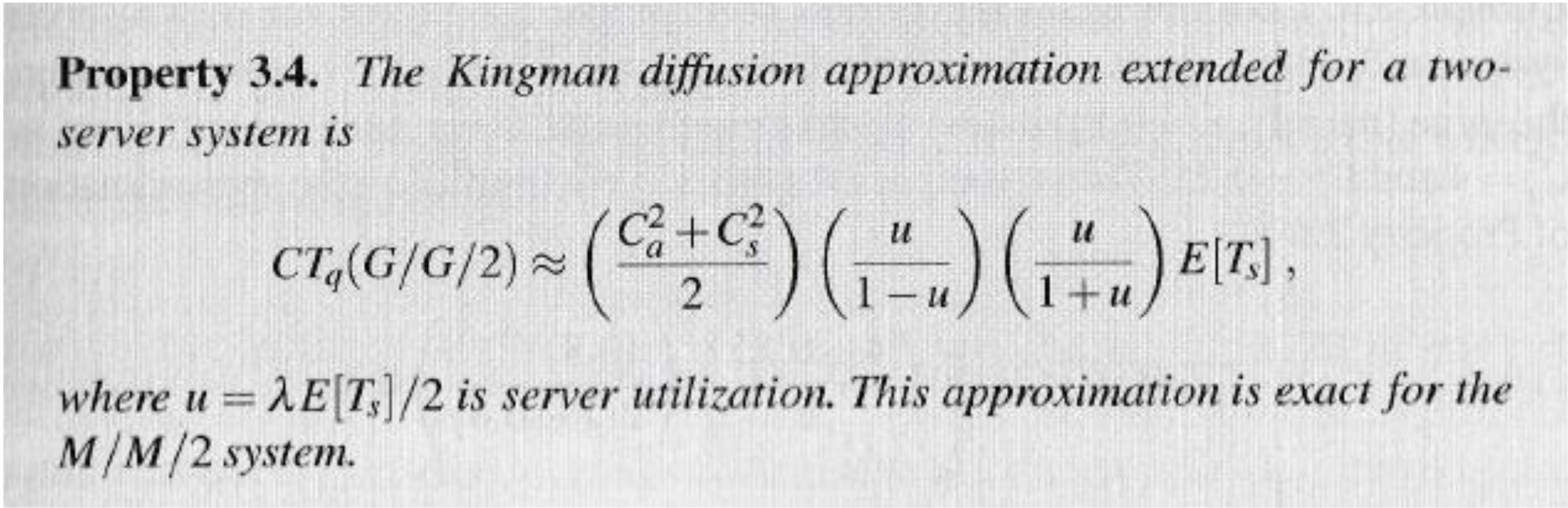


Example 3.7. Consider a $G/G/1$ system with inter-arrival times distributed according to a gamma distribution with mean 15 minutes and standard deviation 30 minutes, and with service times distributed according to an Erlang-4 distribution with mean 12 minutes. Since the distribution of service times is Erlang, the initial temptation may be to use the methodology of Sect. 3.6.1; however, because the arrival times are not exponential, we are left with the $G/G/1$ results. The given data yields the following parameters: $\lambda = 4/\text{hr}$, $\mu = 5/\text{hr}$, $C_a^2 = 4$, and $C_s^2 = 0.25$. Thus, this example has the same mean characteristics of Example 3.6 yielding a utilization of $u = 0.8$, but the arrival process has more variability and the processing times are less variable. Using the Kingman diffusion approximation (Property 3.3), we have

$$CT_q(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right)E[T_s] = \left(\frac{4+0.25}{2}\right)\left(\frac{0.8}{0.2}\right)0.2 = 1.7\,\text{hr}\,.$$

This cycle time is over twice a large as the exponentially distributed system result; thus, the variability associated with non-exponential distributions can have a significant impact on the expected cycle time.

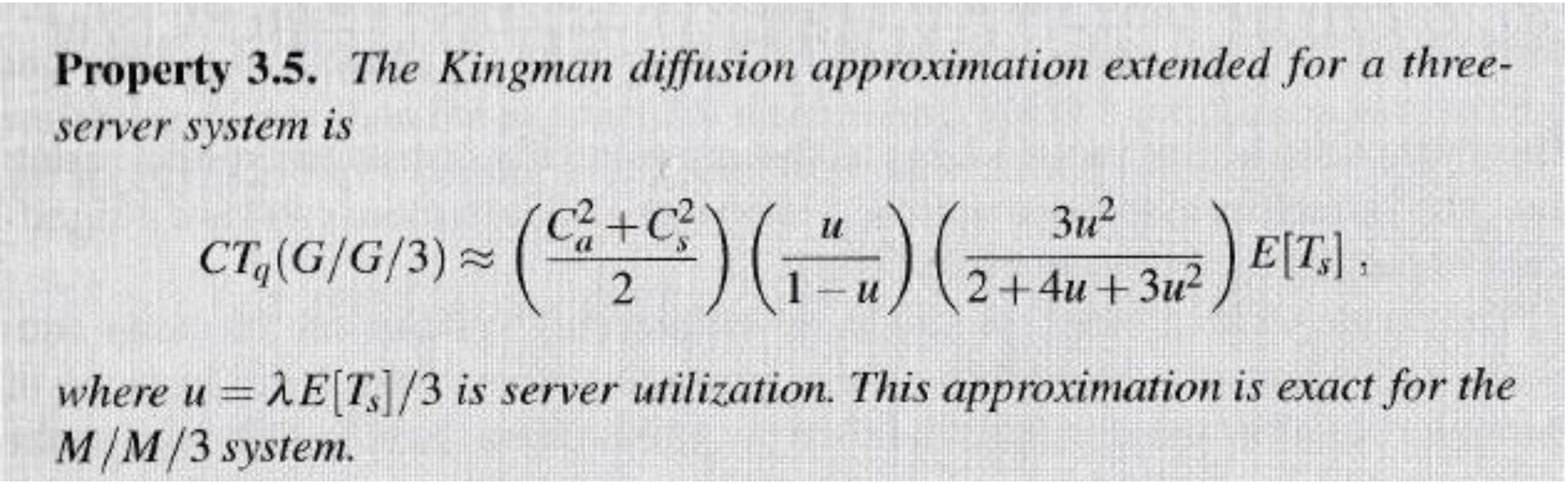
(3) Approximations for G/G/c System



Property 3.4. *The Kingman diffusion approximation extended for a two-server system is*

$$CT_q(G/G/2) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right)\left(\frac{u}{1+u}\right)E[T_s]\,,$$

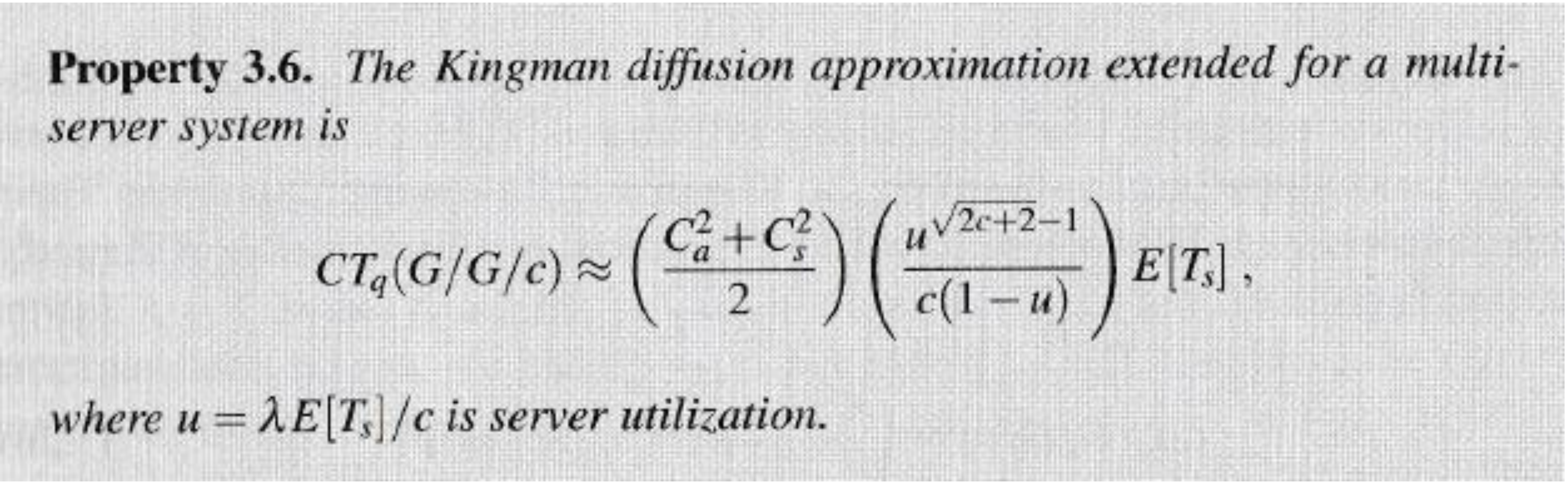
where $u = \lambda E[T_s]/2$ is server utilization. This approximation is exact for the $M/M/2$ system.



Property 3.5. *The Kingman diffusion approximation extended for a three-server system is*

$$CT_q(G/G/3) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right)\left(\frac{3u^2}{2+4u+3u^2}\right)E[T_s]\,,$$

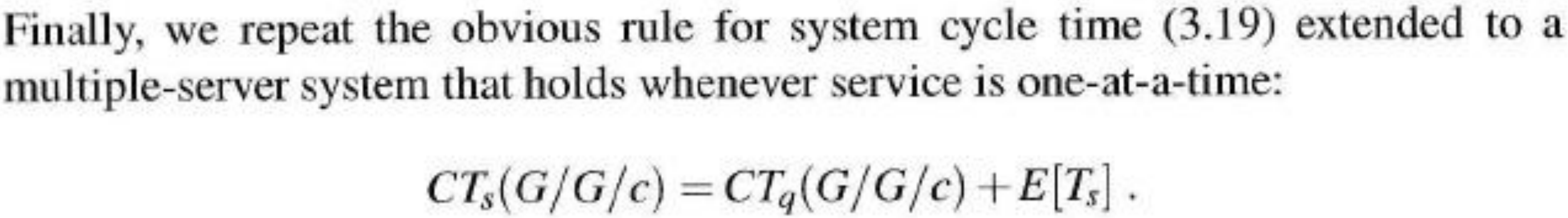
where $u = \lambda E[T_s]/3$ is server utilization. This approximation is exact for the $M/M/3$ system.



Property 3.6. *The Kingman diffusion approximation extended for a multi-server system is*

$$CT_q(G/G/c) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u^{\sqrt{2c+2}-1}}{c(1-u)}\right)E[T_s]\,,$$

where $u = \lambda E[T_s]/c$ is server utilization.



Finally, we repeat the obvious rule for system cycle time (3.19) extended to a multiple-server system that holds whenever service is one-at-a-time:

$$CT_s(G/G/c) = CT_q(G/G/c) + E[T_s]\,.$$

Example 3.8. Consider again the system of Example 3.7 except for a two-server system and with a mean service time of 24 minutes. Thus, server utilization stays the same (namely, $u = 0.8$) and the squared coefficients of variation are still given as $C_a^2 = 4$ and $C_s^2 = 0.25$. Then the expected system cycle time using the approximation of Property 3.6 is

$$CT_q(G/G/2) \approx \left(\frac{4+0.25}{2}\right)\left(\frac{(0.8)^{\sqrt{6}-1}}{2(1-0.8)}\right)0.4$$

$$= 1.54 \text{ hr} .$$

If we use Property 3.4, the approximation becomes

$$CT_q(G/G/2) \approx \left(\frac{4+0.25}{2}\right)\left(\frac{0.8}{1-0.8}\right)\left(\frac{0.8}{1+0.8}\right)0.4$$

$$= 1.51 \text{ hr} .$$

A simulation of this system yielded a mean cycle time in the queue of 1.63 hr with a half-width of ±0.01 hr for the 95% confidence interval. □

1.4 Processing Time Variability

- An approximation for the cycle time in a system queue (or waiting time in the queue for a machine) is denoted as

$$CT_q(G/G/1) = \frac{(C_a^2 + C_s^2)}{2}\left(\frac{u}{1-u}\right)E[T_s]$$

  - Reduce cycle time in the queue by reducing one of the variability components, $C_a{}^2$ or $C_s{}^2$
  - Reducing variability is equivalent to reducing the machine utilization by some factor with respect to the mean cycle time measure.
  - Reducing process variability is equivalent to finding extra capacity in the system since reduction of utilization with a constant arrival rate implies an increase in the mean processing rate.

- To illustrate the equivalence between reducing variability and utilization, consider a single machine system with the following parameter values:

$$C_a^2 = 1$$
$$C_s^2 = 1$$
$$u = 0.8$$
$$E[T_s] = 2 \text{ hr}.$$

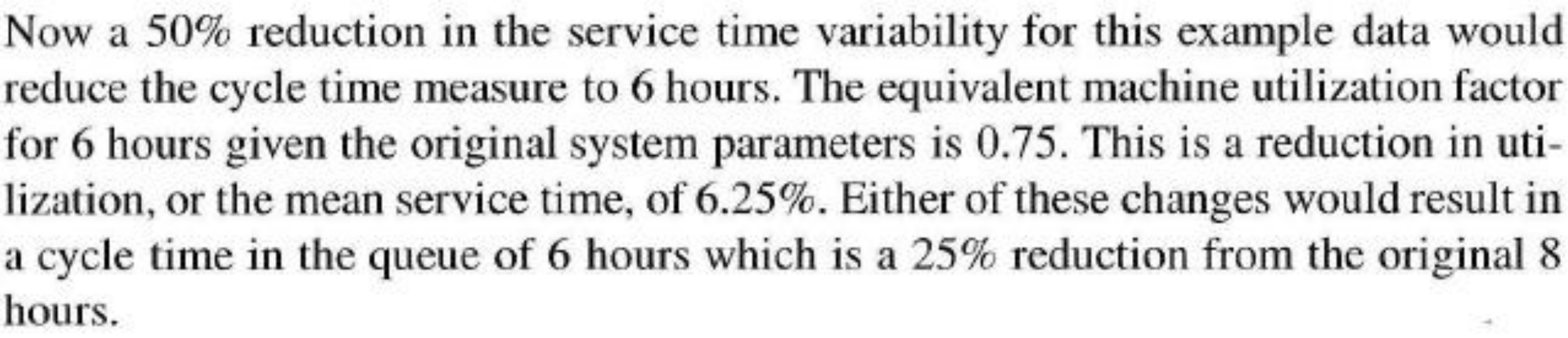
$$CT_q = \frac{(1+1)}{2}\left(\frac{0.8}{1-0.8}\right) 2 \text{ hr} = 8 \text{ hr}.$$

Cycle time in the queue is reduced by 5%.

$$C_s^2 = 0.9$$

$$CT_q = \frac{(1+0.9)}{2}\left(\frac{0.8}{1-0.8}\right) 2 \text{ hr} = 7.6 \text{ hr}$$

$$\text{if } C_s^2 \text{ was not changed } \frac{(1+1)}{2}\left(\frac{u}{1-u}\right) 2 = 7.6,\; u = 0.7917.$$



Now a 50% reduction in the service time variability for this example data would reduce the cycle time measure to 6 hours. The equivalent machine utilization factor for 6 hours given the original system parameters is 0.75. This is a reduction in utilization, or the mean service time, of 6.25%. Either of these changes would result in a cycle time in the queue of 6 hours which is a 25% reduction from the original 8 hours.

- There are many factors that contribute to the variability of the length of the time that a job spends in processing.
  - Natural processing time variability.
  - Random breakdowns and repairs during processing- the variability of the time between breakdowns and the variability of the time to repair a broken machine.
  - Operator unavailability can induce random delays in the time a job spends “in control of ” a machine.
  - Job class setup and take-down times- the time caused by a job-type change on a machine.

1.4.1 Natural Processing Time Variability

- Consider a job with processing time random variable, T, with known mean and variance parameters $E[T]$ and $V[T]$, respectively. It is assumed that T is made up of three separate (independent) sub-tasks. Hence,

$$E[T] = E[T_1] + E[T_2] + E[T_3]$$

$$V[T] = V[T_1] + V[T_2] + V[T_3]$$

$$C^2[T] = \frac{V[T]}{E[T]^2} .$$

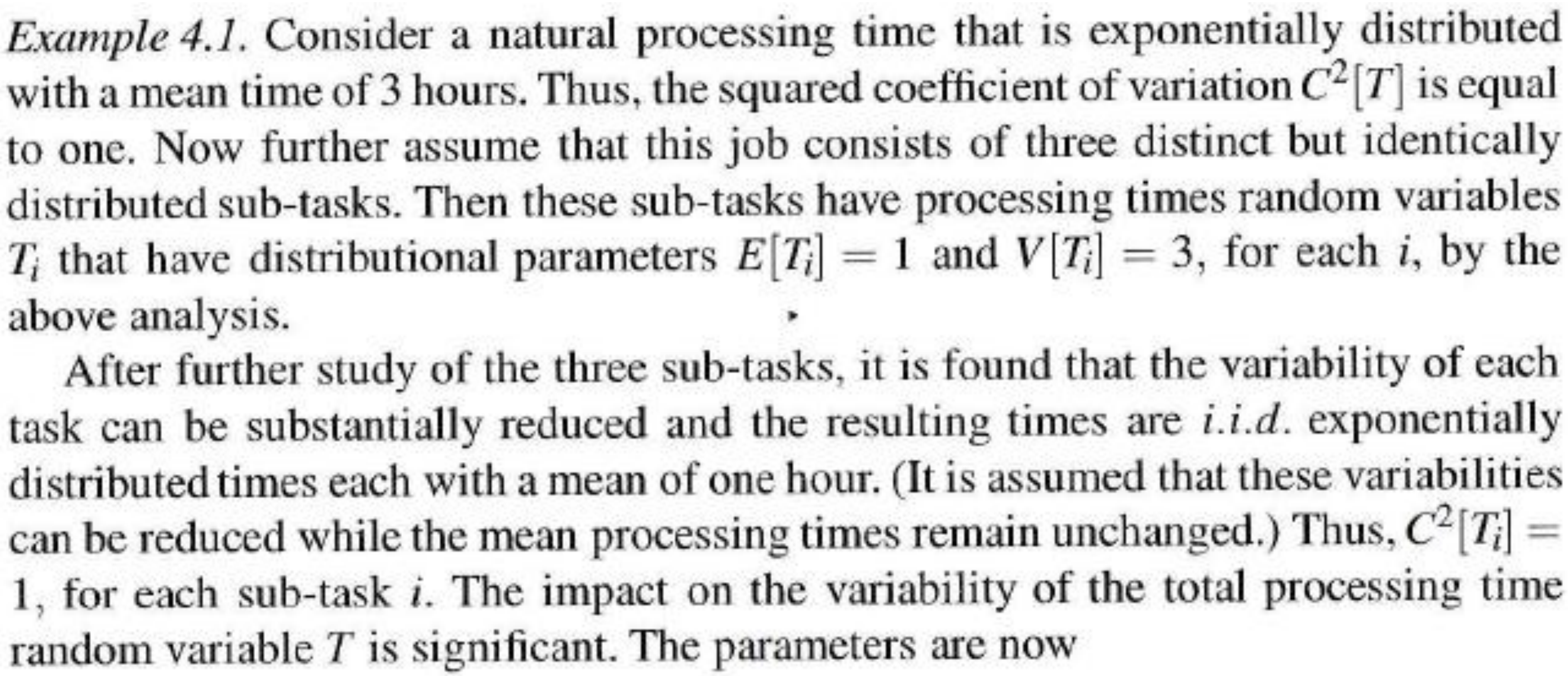
Additionally consider that these three sub-processes times are independent and identically distributed random variables so that

$$E[T] = 3E[T_1] \qquad V[T] = 3V[T_1]$$

→

$$E[T_i] = \frac{E[T]}{3} \qquad V[T_i] = \frac{V[T]}{3}$$

$$C^2[T_i] = \frac{V[T_i]}{E[T_i]^2} = \frac{V[T]/3}{E[T]^2/3^2} = 3C^2[T], \quad \text{for } i = 1, 2, 3 .$$



Example 4.1. Consider a natural processing time that is exponentially distributed with a mean time of 3 hours. Thus, the squared coefficient of variation $C^2[T]$ is equal to one. Now further assume that this job consists of three distinct but identically distributed sub-tasks. Then these sub-tasks have processing times random variables T_i that have distributional parameters $E[T_i] = 1$ and $V[T_i] = 3$, for each i, by the above analysis.

After further study of the three sub-tasks, it is found that the variability of each task can be substantially reduced and the resulting times are *i.i.d.* exponentially distributed times each with a mean of one hour. (It is assumed that these variabilities can be reduced while the mean processing times remain unchanged.) Thus, $C^2[T_i] = 1$, for each sub-task i. The impact on the variability of the total processing time random variable T is significant. The parameters are now

$$E[T_i] = 1$$

$$V[T_i] = 1$$

$$C^2[T_i] = 1$$

$$E[X] = \frac{1}{\lambda};\ V[X] = \frac{1}{\lambda^2};\ C^2[X] = \frac{V[X]}{E[X]^2} = 1\,.$$

The total processing time variability was reduced to 1/3 of its original value, which reduced the associated workstation cycle time in the queue. Extra processing capability has found with a faster processing time.

$$E[T] = \sum_{i=1}^{3} E[T_i] = 3$$

$$V[T] = \sum_{i=1}^{3} V[T_i] = 3$$

$$C^2[T] = \frac{3}{3^2} = 1/3\,.$$

1.4.2 Random Breakdowns and Repairs During Processing

- Several courses of action might result from the breakdown of a machine.
  - The job undergoing processing at the time of breakdown might not be recoverable (i.e., lost)
  - The job might require additional processing before resumption of “normal” processing
  - The job might not be effected by the breakdown and normal processing can resume immediately after the repair is complete (as if the breakdown never occurred).
- Here we consider the two latter situations (for the second case the additional processing time need to resume service is included in the machine repair time).

- *Definition*. The *effective processing time*, T_e, refers to the time that a job first has control of the processor until the time at which the job releases the processor so that it is available to begin work on another job.

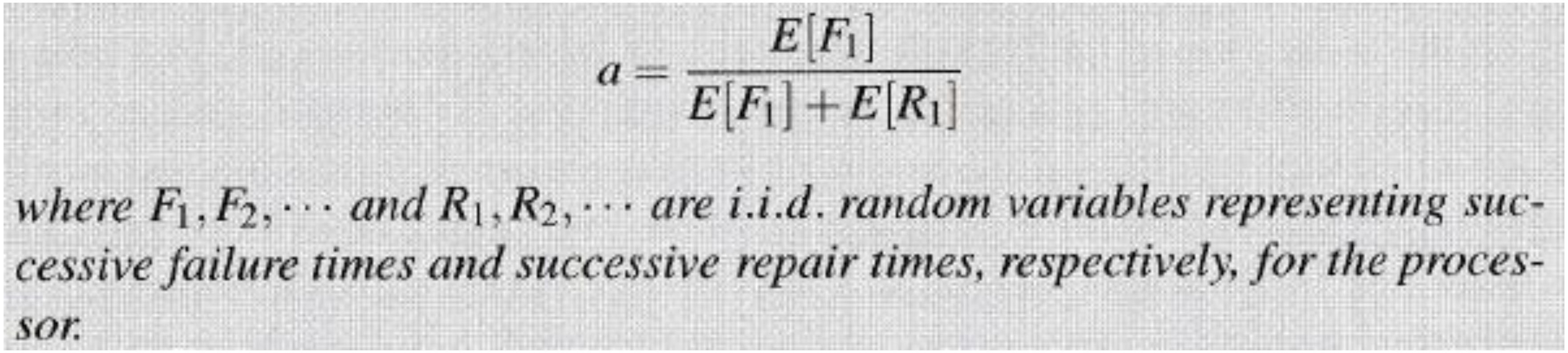
$$T_e = T + \sum_{i=1}^{N} R_i$$

T: the normal (uninterrupted) processing time random variable

R_i: the repair time random variables

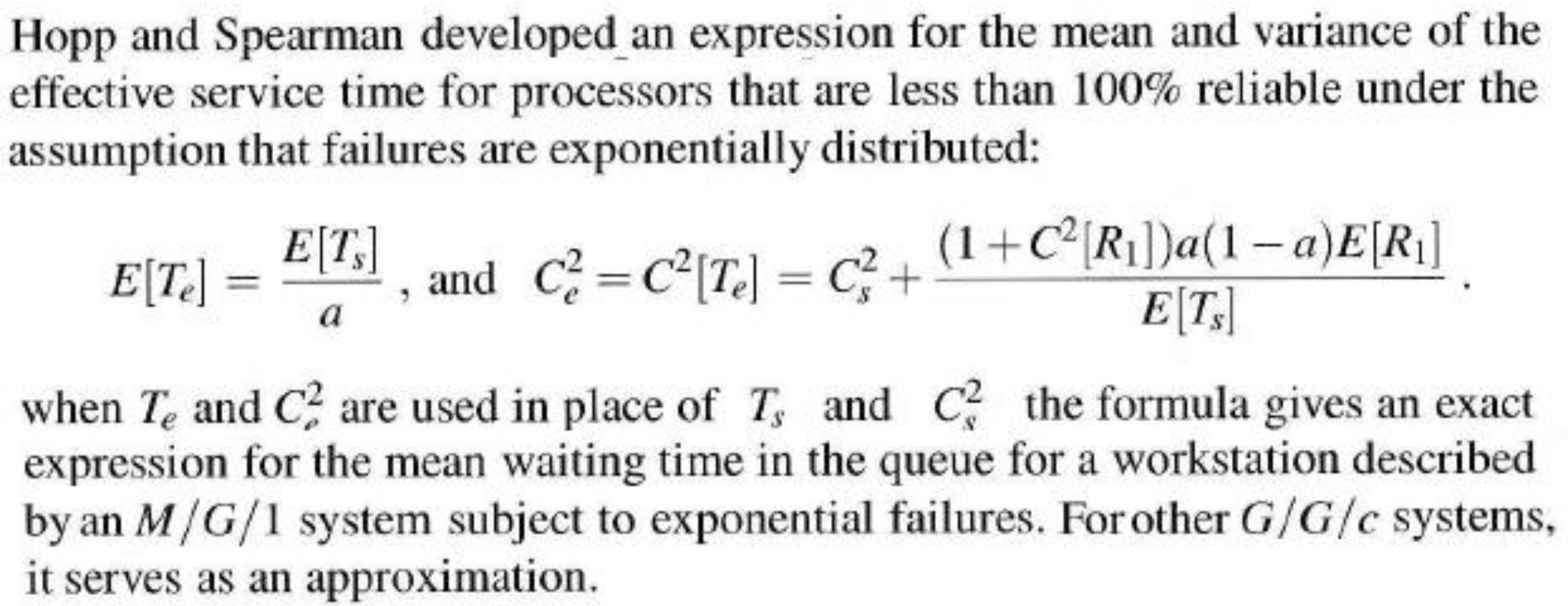
N: the random number of failures during the service time T

- *Definition*. The *availability*, *a*, of a processor that is subject to failures is the long-run average fraction of time that the processor is available for processing jobs. *Processor availability* is determined by



$$a = \frac{E[F_1]}{E[F_1]+E[R_1]}$$

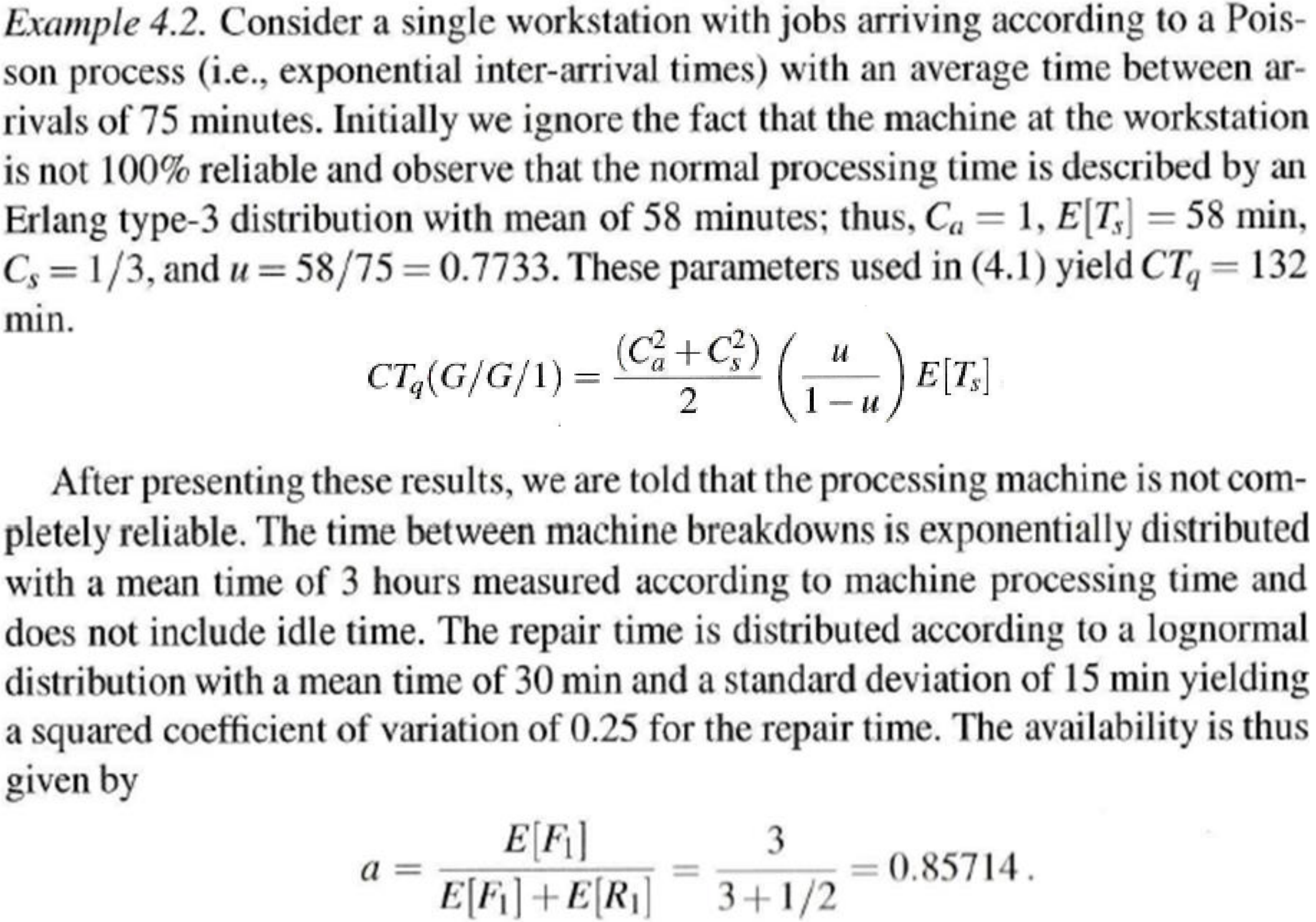
where $F_1, F_2, \cdots$ and $R_1, R_2, \cdots$ are i.i.d. random variables representing successive failure times and successive repair times, respectively, for the processor.



Hopp and Spearman developed an expression for the mean and variance of the effective service time for processors that are less than 100% reliable under the assumption that failures are exponentially distributed:

$$E[T_e] = \frac{E[T_s]}{a}, \text{ and } \quad C_e^2 = C^2[T_e] = C_s^2 + \frac{(1+C^2[R_1])a(1-a)E[R_1]}{E[T_s]}.$$

when T_e and C_e^2 are used in place of T_s and C_s^2 the formula gives an exact expression for the mean waiting time in the queue for a workstation described by an $M/G/1$ system subject to exponential failures. For other $G/G/c$ systems, it serves as an approximation.

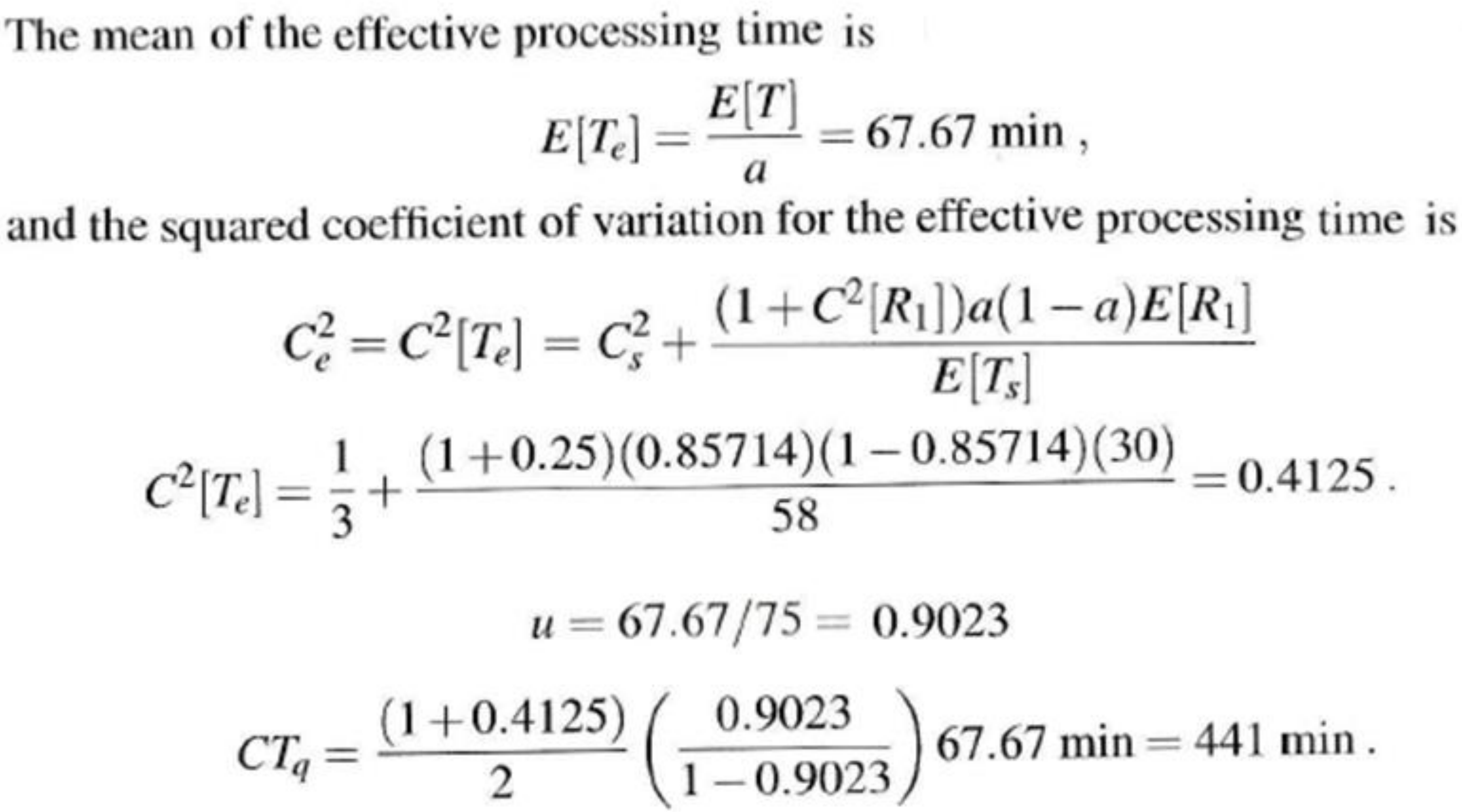


Example 4.2. Consider a single workstation with jobs arriving according to a Poisson process (i.e., exponential inter-arrival times) with an average time between arrivals of 75 minutes. Initially we ignore the fact that the machine at the workstation is not 100% reliable and observe that the normal processing time is described by an Erlang type-3 distribution with mean of 58 minutes; thus, $C_a = 1$, $E[T_s] = 58$ min, $C_s = 1/3$, and $u = 58/75 = 0.7733$. These parameters used in (4.1) yield $CT_q = 132$ min.

$$CT_q(G/G/1) = \frac{(C_a^2 + C_s^2)}{2}\left(\frac{u}{1-u}\right)E[T_s]$$

After presenting these results, we are told that the processing machine is not completely reliable. The time between machine breakdowns is exponentially distributed with a mean time of 3 hours measured according to machine processing time and does not include idle time. The repair time is distributed according to a lognormal distribution with a mean time of 30 min and a standard deviation of 15 min yielding a squared coefficient of variation of 0.25 for the repair time. The availability is thus given by

$$a = \frac{E[F_1]}{E[F_1] + E[R_1]} = \frac{3}{3 + 1/2} = 0.85714\,.$$



The mean of the effective processing time is

$$E[T_e] = \frac{E[T]}{a} = 67.67 \text{ min},$$

and the squared coefficient of variation for the effective processing time is

$$C_e^2 = C^2[T_e] = C_s^2 + \frac{(1 + C^2[R_1])a(1-a)E[R_1]}{E[T_s]}$$

$$C^2[T_e] = \frac{1}{3} + \frac{(1+0.25)(0.85714)(1-0.85714)(30)}{58} = 0.4125 .$$

$$u = 67.67/75 = 0.9023$$

$$CT_q = \frac{(1+0.4125)}{2}\left(\frac{0.9023}{1-0.9023}\right) 67.67 \text{ min} = 441 \text{ min} .$$

- The inclusion of machine failure in the model results in over a three-fold increase in the mean waiting time.
  - Machine failures cause an increase the effective utilization factor. As the utilization factor approaches one, small changes in the factor will have major changes in waiting times.
  - Machine failures cause an increase in the service variability.

1.4.3 Operator Variability

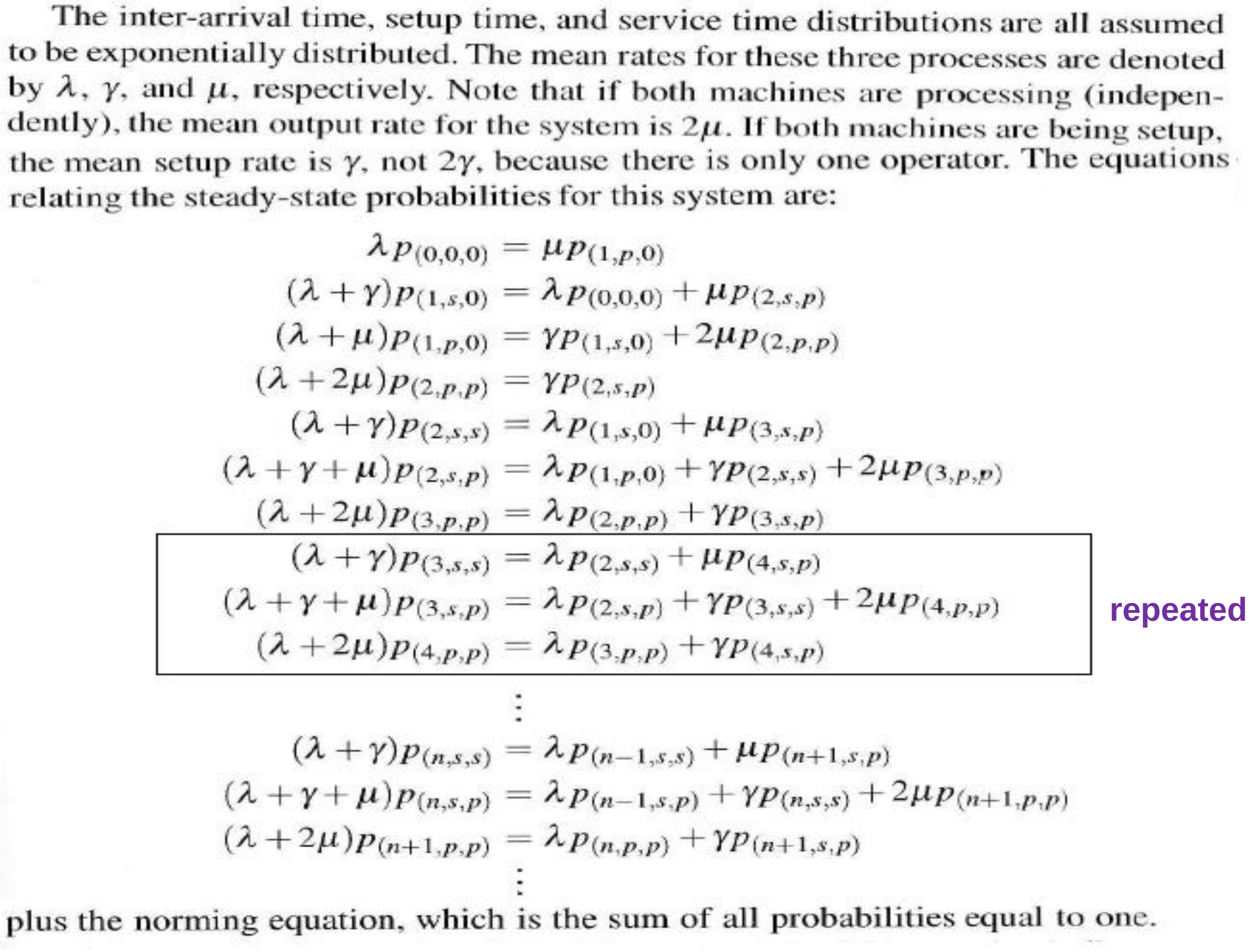
- Operators are frequently required to setup a machine for each job.
- If an operator is assigned to cover too many machines then system performance can be significantly degraded because of delays resulting from waiting for the operator to become available to perform the necessary job setups.
- If a system has reasonable capacity, then the operator machine interaction problem does not significantly impact system performance. Thus, this level of detail is frequently omitted in system models.

It is assumed that one job class is treated with two identical machines and one operator. A three-tuple (n, i, j) is used to represent the state of the system, where n denotes the number of jobs in the system and i and j indicate the status of the two machines. There are three possible values for i and j: 0 indicates a machine has no job associated with it, s indicates that a machine has a job "in setup", and p indicates a machine has a job "in process".

For example:

- State (1, *s*, 0): there is one job in the system and the operator is setting it up on a machine.
- State (5, *s*, *s*): there are 5 jobs in the system with one job being set-up on a machine, another job waiting at a machine for the operator, and 3 jobs waiting in the queue for a machine.
- State (7, *p*, *p*): there are 7 jobs in the system with both machines busy processing, 5 jobs queued, and the operator idle.
- The state space representation for *n*≥2 is made up of three individual states: (*n*, *s*, *s*), (*n*, *s*, *p*), and (*n*, *p*, *p*).

$$\{(0,0,0),(1,s,0),(1,p,0),(2,s,s),(2,s,p),(2,p,p),(3,s,s),(3,s,p),(3,p,p),\cdots\}$$



The inter-arrival time, setup time, and service time distributions are all assumed to be exponentially distributed. The mean rates for these three processes are denoted by λ, γ, and μ, respectively. Note that if both machines are processing (independently), the mean output rate for the system is 2μ. If both machines are being setup, the mean setup rate is γ, not 2γ, because there is only one operator. The equations relating the steady-state probabilities for this system are:

$$\begin{aligned}
\lambda p_{(0,0,0)} &= \mu p_{(1,p,0)} \\
(\lambda+\gamma) p_{(1,s,0)} &= \lambda p_{(0,0,0)} + \mu p_{(2,s,p)} \\
(\lambda+\mu) p_{(1,p,0)} &= \gamma p_{(1,s,0)} + 2\mu p_{(2,p,p)} \\
(\lambda+2\mu) p_{(2,p,p)} &= \gamma p_{(2,s,p)} \\
(\lambda+\gamma) p_{(2,s,s)} &= \lambda p_{(1,s,0)} + \mu p_{(3,s,p)} \\
(\lambda+\gamma+\mu) p_{(2,s,p)} &= \lambda p_{(1,p,0)} + \gamma p_{(2,s,s)} + 2\mu p_{(3,p,p)} \\
(\lambda+2\mu) p_{(3,p,p)} &= \lambda p_{(2,p,p)} + \gamma p_{(3,s,p)}
\end{aligned}$$

$$\boxed{\begin{aligned}
(\lambda+\gamma) p_{(3,s,s)} &= \lambda p_{(2,s,s)} + \mu p_{(4,s,p)} \\
(\lambda+\gamma+\mu) p_{(3,s,p)} &= \lambda p_{(2,s,p)} + \gamma p_{(3,s,s)} + 2\mu p_{(4,p,p)} \\
(\lambda+2\mu) p_{(4,p,p)} &= \lambda p_{(3,p,p)} + \gamma p_{(4,s,p)}
\end{aligned}} \quad \textbf{repeated}$$

$$\begin{aligned}
&\vdots \\
(\lambda+\gamma) p_{(n,s,s)} &= \lambda p_{(n-1,s,s)} + \mu p_{(n+1,s,p)} \\
(\lambda+\gamma+\mu) p_{(n,s,p)} &= \lambda p_{(n-1,s,p)} + \gamma p_{(n,s,s)} + 2\mu p_{(n+1,p,p)} \\
(\lambda+2\mu) p_{(n+1,p,p)} &= \lambda p_{(n,p,p)} + \gamma p_{(n+1,s,p)} \\
&\vdots
\end{aligned}$$

plus the norming equation, which is the sum of all probabilities equal to one.

To be more specific, we first observe that $p_{(1,p,0)} = (\lambda/\mu)p_{(0,0,0)}$. The the second through fourth equations can be rewritten in matrix form as

$$\begin{bmatrix} -(\lambda+\gamma) & \mu & 0 \\ \gamma & 0 & 2\mu \\ 0 & \gamma & -(\lambda+2\mu) \end{bmatrix} \begin{bmatrix} p_{(1,s,0)} \\ p_{(2,s,p)} \\ p_{(2,p,p)} \end{bmatrix} = \begin{bmatrix} -\lambda p_{(0,0,0)} \\ (\lambda+\mu)p_{(1,p,0)} \\ 0 \end{bmatrix},$$

with its solution given by

$$\begin{bmatrix} p_{(1,s,0)} \\ p_{(2,s,p)} \\ p_{(2,p,p)} \end{bmatrix} = \begin{bmatrix} -(\lambda+\gamma) & \mu & 0 \\ \gamma & 0 & 2\mu \\ 0 & \gamma & -(\lambda+2\mu) \end{bmatrix}^{-1} \begin{bmatrix} -\lambda p_{(0,0,0)} \\ (\lambda+\mu)p_{(1,p,0)} \\ 0 \end{bmatrix}.$$

Once the values of the probabilities $(p_{(1,s,0)}, p_{(2,s,p)}, p_{(2,p,p)})$ have been obtained, the vector $(p_{(2,s,s)}, p_{(3,s,p)}, p_{(3,p,p)})$ is solved similarly using the fifth through seventh equations in the system. This solution is written as

$$\begin{bmatrix} p_{(2,s,s)} \\ p_{(3,s,p)} \\ p_{(3,p,p)} \end{bmatrix} = \begin{bmatrix} -(\lambda+\gamma) & \mu & 0 \\ \gamma & 0 & 2\mu \\ 0 & \gamma & -(\lambda+2\mu) \end{bmatrix}^{-1} \begin{bmatrix} -\lambda p_{(1,s,0)} \\ \lambda p_{(1,p,0)} - (\lambda+\mu+\gamma)p_{(2,s,p)} \\ \lambda p_{(2,p,p)} \end{bmatrix}.$$

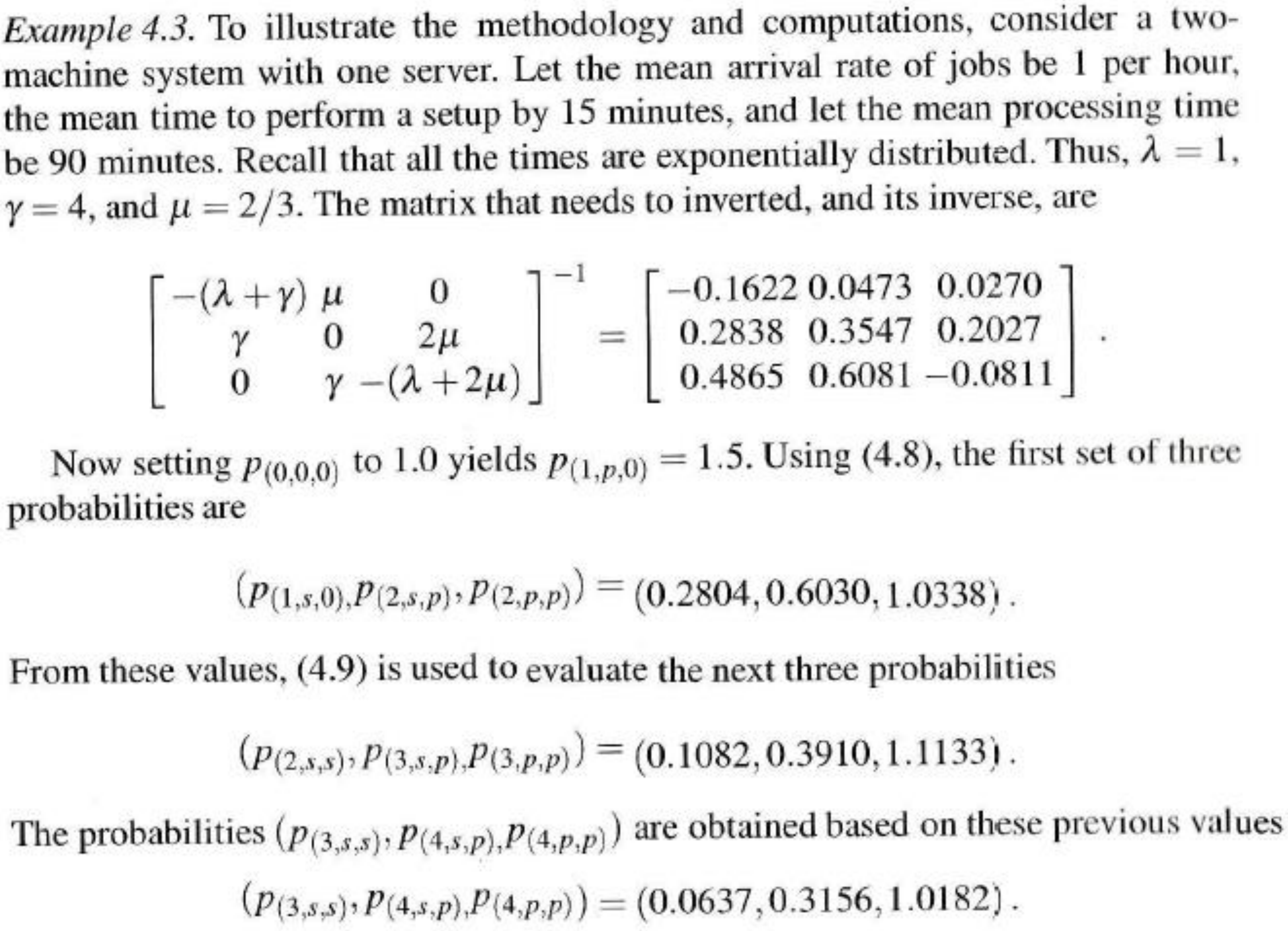
plus the norming equation, which is the sum of all probabilities equal to one.

$$\begin{bmatrix} p_{(n,s,s)} \\ p_{(n+1,s,p)} \\ p_{(n+1,p,p)} \end{bmatrix} = \boxed{\begin{bmatrix} -(\lambda+\gamma) & \mu & 0 \\ \gamma & 0 & 2\mu \\ 0 & \gamma & -(\lambda+2\mu) \end{bmatrix}^{-1}} \begin{bmatrix} -\lambda p_{(n-1,s,s)} \\ \lambda p_{(n-1,s,p)} - (\lambda+\mu+\gamma) p_{(n,s,p)} \\ \lambda p_{(n,p,p)} \end{bmatrix}$$

Notice that the solution to each system always involves the same inverse which greatly simplifies the computational burden of the process.

- If the operator sets up too slowly or if the arrival rates are too fast for the processing times, the queues will build up continually and no steady-state is possible. Steady-state probabilities will exist if and only if the three parameter values are such that

$$\frac{2(\mu+\gamma)\mu\gamma}{2\mu^2+2\mu\gamma+\gamma^2} < \lambda$$



Example 4.3. To illustrate the methodology and computations, consider a two-machine system with one server. Let the mean arrival rate of jobs be 1 per hour, the mean time to perform a setup by 15 minutes, and let the mean processing time be 90 minutes. Recall that all the times are exponentially distributed. Thus, $\lambda = 1$, $\gamma = 4$, and $\mu = 2/3$. The matrix that needs to inverted, and its inverse, are

$$\begin{bmatrix} -(\lambda+\gamma) & \mu & 0 \\ \gamma & 0 & 2\mu \\ 0 & \gamma & -(\lambda+2\mu) \end{bmatrix}^{-1} = \begin{bmatrix} -0.1622 & 0.0473 & 0.0270 \\ 0.2838 & 0.3547 & 0.2027 \\ 0.4865 & 0.6081 & -0.0811 \end{bmatrix}.$$

Now setting $p_{(0,0,0)}$ to 1.0 yields $p_{(1,p,0)} = 1.5$. Using (4.8), the first set of three probabilities are

$$\left(p_{(1,s,0)}, p_{(2,s,p)}, p_{(2,p,p)}\right) = (0.2804, 0.6030, 1.0338)\,.$$

From these values, (4.9) is used to evaluate the next three probabilities

$$\left(p_{(2,s,s)}, p_{(3,s,p)}, p_{(3,p,p)}\right) = (0.1082, 0.3910, 1.1133)\,.$$

The probabilities $\left(p_{(3,s,s)}, p_{(4,s,p)}, p_{(4,p,p)}\right)$ are obtained based on these previous values

$$\left(p_{(3,s,s)}, p_{(4,s,p)}, p_{(4,p,p)}\right) = (0.0637, 0.3156, 1.0182)\,.$$

Repeating obtain

$$\left(p_{(4,s,s)}, p_{(5,s,p)}, p_{(5,p,p)}\right) = (0.0489, 0.2713, 0.9014),$$
$$\left(p_{(5,s,s)}, p_{(6,s,p)}, p_{(6,p,p)}\right) = (0.0413, 0.2367, 0.7921),$$
$$\vdots$$
$$\left(p_{(14,s,s)}, p_{(15,s,p)}, p_{(15,p,p)}\right) = (0.0110, 0.0635, 0.2129).$$

Stopping at this point, these probabilities sum to 15.288. Dividing all of these probabilities by 15.288 yields an approximate solution to this system. It is obvious that since the probability $p_{(15,p,p)}$ is not very close to zero, that this truncated solution will not be very close to the unlimited system solution. In fact using these probability values, the estimate for the mean number of jobs, N_s, in the system is

$$WIP = E[N_s] = 5.606\,.$$

As the number of probabilities obtained is increased, expected system WIP, converges.

$$n = 20,\ WIP = 6.399,$$
$$n = 30,\ WIP = 7.263,$$
$$n = 40,\ WIP = 7.603,$$
$$n = 50,\ WIP = 7.725,$$
$$n = 60,\ WIP = 7.658,$$
$$n = 70,\ WIP = 7.779,$$
$$n = 80,\ WIP = 7.783,$$
$$n = 90,\ WIP = 7.785,$$
$$n = 100,\ WIP = 7.785.$$

The truncated system solution changes very little as more probabilities are added beyond the first 80 probabilities. Thus, a reasonable solution to the unlimited system has been obtained. The expected cycle time in the system from Little's Law is

$$CT = WIP/\lambda = 7.785 \text{ hr} .$$

The expected number of jobs in the operator system is

$$1 \times \left(p_{(1,s,0)} + \sum_{n=2}^{\infty} p_{(n,s,p)} \right) + 2 \times \sum_{n=2}^{\infty} p_{(n,s,s)} = 0.2819 ,$$

with the probability that the operator is idle being

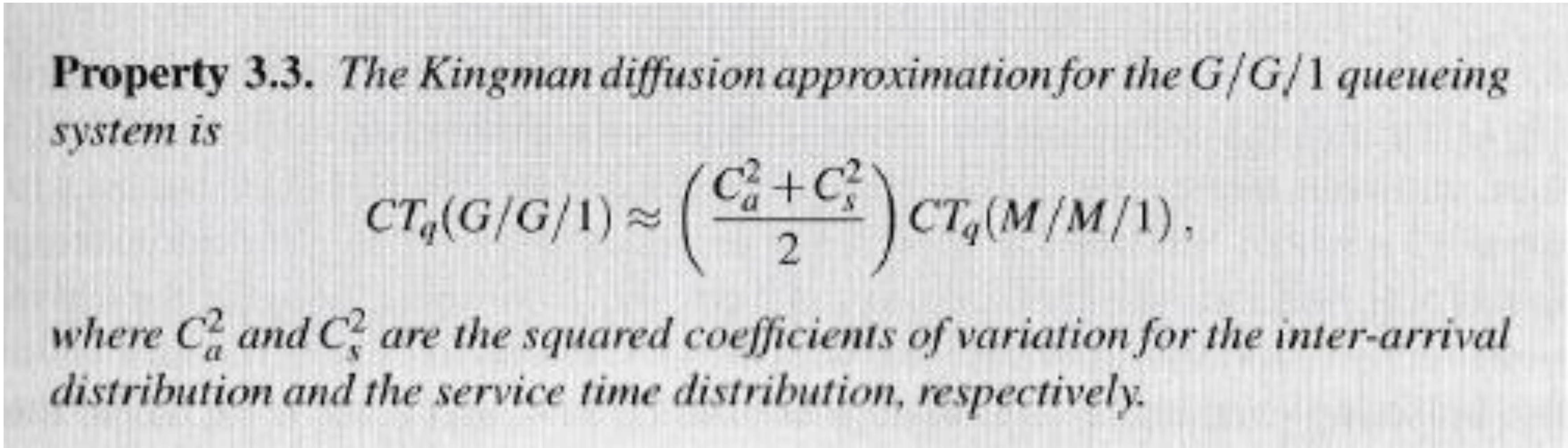
$$p_{(0,0,0)} + p_{(1,p,0)} + \sum_{n=2}^{\infty} p_{(n,p,p)} = 0.75 ,$$

and the machine utilization factor being

$$\frac{1}{2} \times \left(p_{(1,p,0)} + \sum_{n=2}^{\infty} p_{(n,s,p)} \right) + 1 \times \sum_{n=2}^{\infty} p_{(n,p,p)} = 0.8909 .$$

1.5 Multi-Stage Single-Product Factory Models

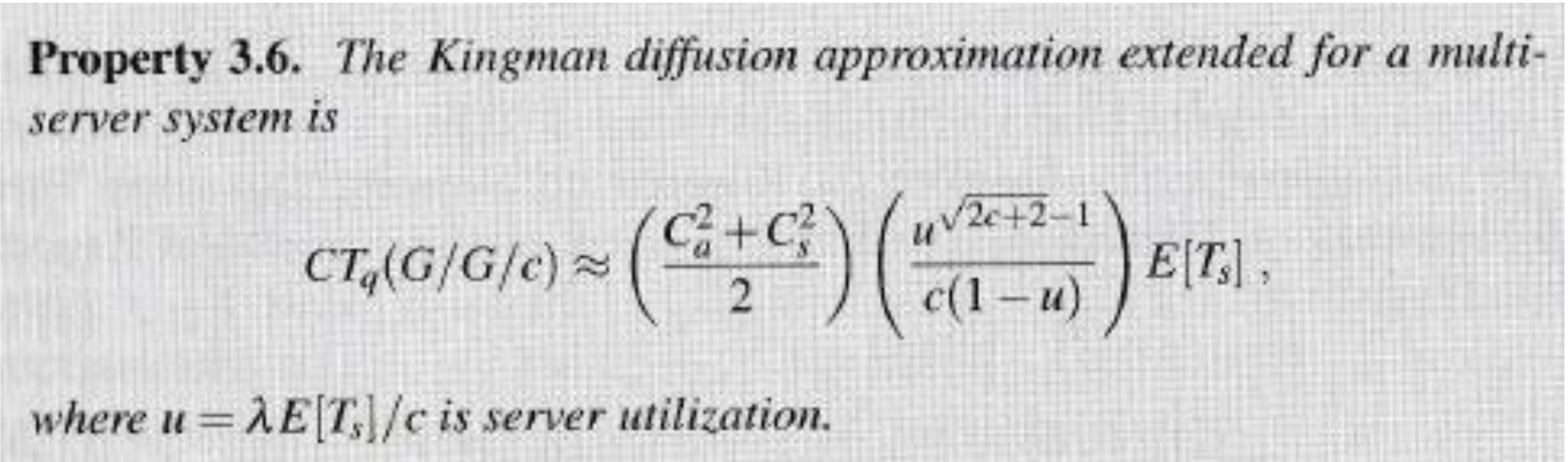
- Linking several workstations together is necessary step towards more realistic factory models.



Property 3.3. *The Kingman diffusion approximation for the G/G/1 queueing system is*

$$CT_q(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2}\right) CT_q(M/M/1),$$

where C_a^2 and C_s^2 are the squared coefficients of variation for the inter-arrival distribution and the service time distribution, respectively.



Property 3.6. *The Kingman diffusion approximation extended for a multi-server system is*

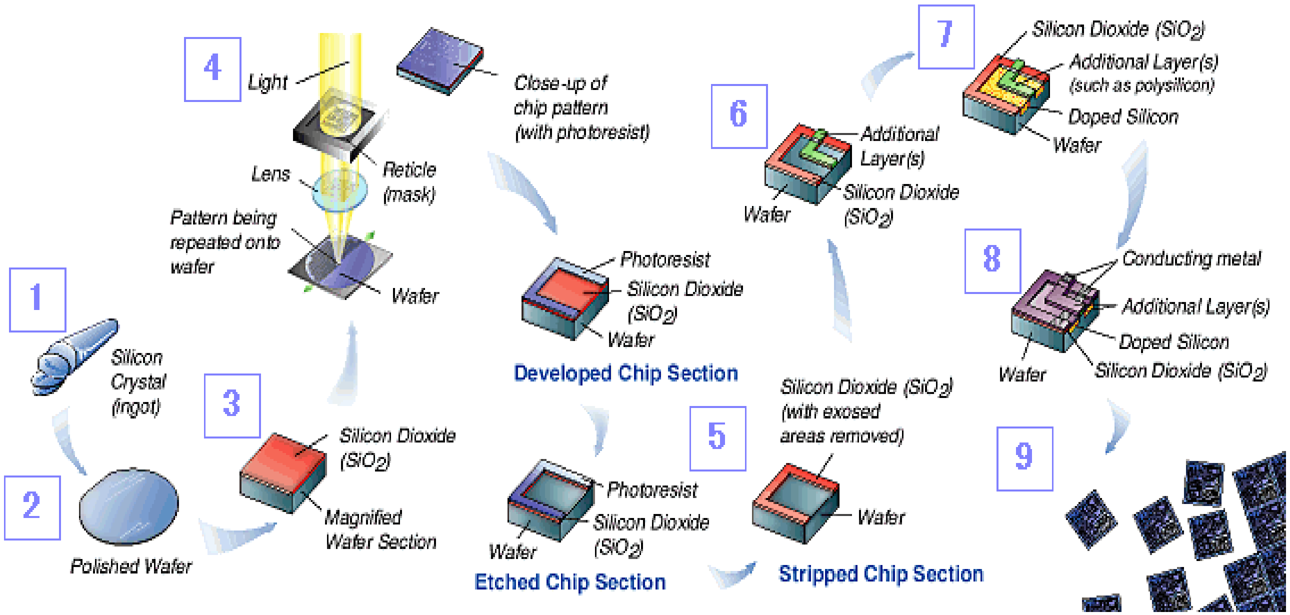
$$CT_q(G/G/c) \approx \left(\frac{C_a^2 + C_s^2}{2}\right) \left(\frac{u^{\sqrt{2c+2}-1}}{c(1-u)}\right) E[T_s],$$

where $u = \lambda E[T_s]/c$ is server utilization.

- For general system configurations, there are two basic mechanisms that must be explored:

  (1) The merging of several input streams into a workstation.

  (2) The separation or partitioning of a workstation output stream into several different streams for different target workstations.

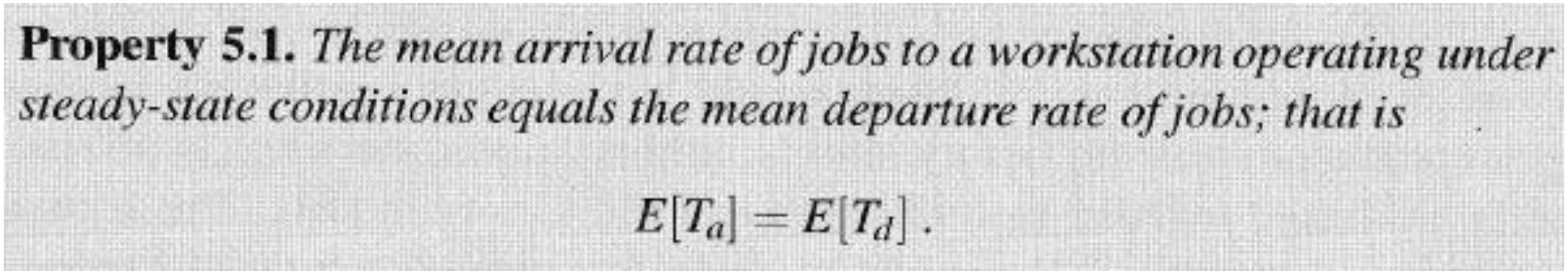


1
Silicon Crystal (ingot)
2
Polished Wafer
3
Silicon Dioxide (SiO_2)
Magnified Wafer Section
4
Light
Reticle (mask)
Lens
Pattern being repeated onto wafer
Wafer
Close-up of chip pattern (with photoresist)
Photoresist
Silicon Dioxide (SiO_2)
Wafer
Developed Chip Section
Photoresist
Silicon Dioxide (SiO_2)
Wafer
Etched Chip Section
5
Silicon Dioxide (SiO_2) (with exosed areas removed)
Wafer
Stripped Chip Section
6
Additional Layer(s)
Silicon Dioxide (SiO_2)
Wafer
7
Silicon Dioxide (SiO_2)
Additional Layer(s) (such as polysilicon)
Doped Silicon
Wafer
8
Conducting metal
Additional Layer(s)
Doped Silicon
Silicon Dioxide (SiO_2)
Wafer
9

- We start with workstations in series and progress to more complex general network configurations.

1.5.1 Approximation the Departure Process from a Workstation

- How the workstation transforms the inter-arrival process characteristics into output-stream characteristics?



Property 5.1. *The mean arrival rate of jobs to a workstation operating under steady-state conditions equals the mean departure rate of jobs; that is*

$$E[T_a] = E[T_d] .$$

- For M/M/c systems with c≥1, the output process is probabilistically identical to the input process; namely, the inter-departure times are exponentially distributed so that C_d^2=C_a^2=C_s^2.
- For non-exponential systems
  - If the workstation is extremely busy, C_d^2 would be expected to very close in value to C_s^2 .
  - If the system is very lightly loaded, C_d^2 would be expected to very close in to C^2

- For an M/G/1 system, an conclusion proposed by Buzacott and Shanthikumar[3] is exactly the correct.

$$C_d^2(M/G/1) = 1 - u^2 + u^2 C_s^2$$

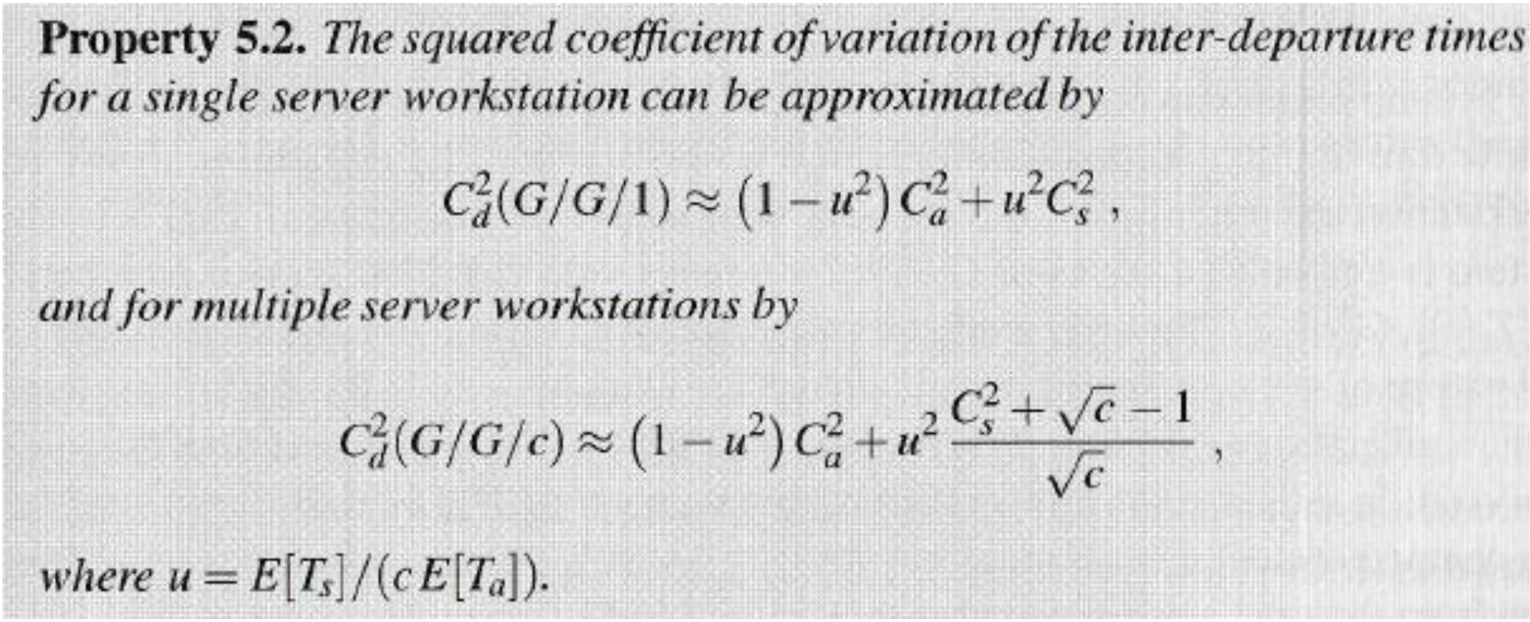
- They also develop for the G/G/1 system a lower bound on C_d^2 as

$$C_d^2(G/G/1) \geq (1-u)\left(1+uC_a^2\right) C_a^2 + u^2 C_s^2$$

- A general relationship for a G/G/1 system for the squared coefficient of variation was developed by Marshall [4] as

$$C_d^2 = C_a^2 + 2u^2 C_s^2 - 2u(1-u)CT_q/E[T_s]$$

$$CT_q = ((C_a^2 + C_s^2)/2)uE[T_s]/(1-u)$$



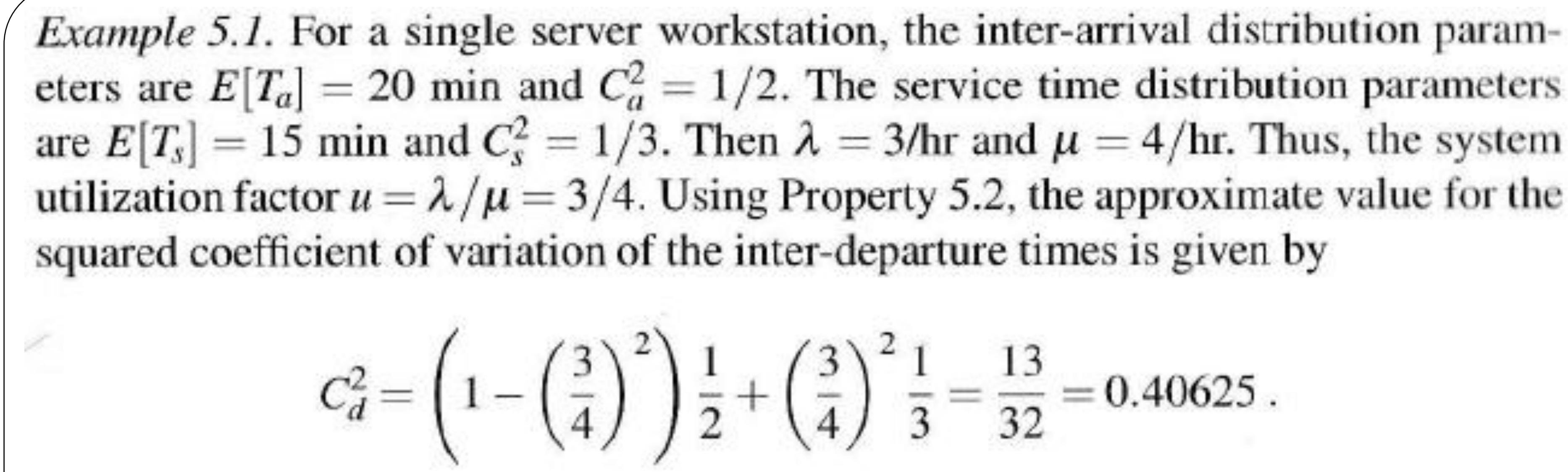
Property 5.2. *The squared coefficient of variation of the inter-departure times for a single server workstation can be approximated by*

$$C_d^2(G/G/1) \approx \left(1-u^2\right) C_a^2 + u^2 C_s^2 \,,$$

and for multiple server workstations by

$$C_d^2(G/G/c) \approx \left(1-u^2\right) C_a^2 + u^2 \frac{C_s^2 + \sqrt{c} - 1}{\sqrt{c}} \,,$$

where $u = E[T_s]/(cE[T_a])$.

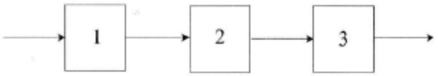


Example 5.1. For a single server workstation, the inter-arrival distribution parameters are $E[T_a] = 20$ min and $C_a^2 = 1/2$. The service time distribution parameters are $E[T_s] = 15$ min and $C_s^2 = 1/3$. Then $\lambda = 3/\text{hr}$ and $\mu = 4/\text{hr}$. Thus, the system utilization factor $u = \lambda/\mu = 3/4$. Using Property 5.2, the approximate value for the squared coefficient of variation of the inter-departure times is given by

$$C_d^2 = \left(1 - \left(\frac{3}{4}\right)^2\right)\frac{1}{2} + \left(\frac{3}{4}\right)^2\frac{1}{3} = \frac{13}{32} = 0.40625 .$$

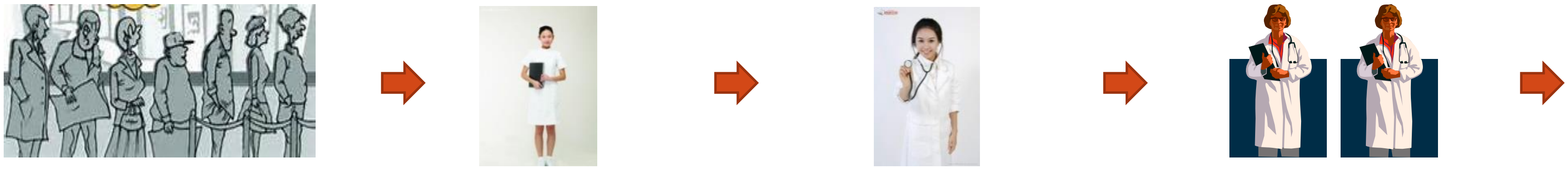
1.5.2 Serial Systems Decomposition

- Consider a pure serial system with external inflow into the first workstation only and no branc



  - The departures from each workstation are the inflows into the next workstation.
  - The system can be treated as a series of G/G/c/∞ queues with specified service parameters ($E[T_s(i)]$, $C_s^2(i)$, c_i) for each workstation *i*, numbered from 1 to *n*.
  - The arrival stream for workstation *i* is the departure stream from workstation *i* -1, i.e., $C_a^2(i)= C_d^2(i\text{-}1)$.

- Burke[2] proved that the output for any M/M/c/∞ system is a Poisson process with the same parameters as the input process but statistically independent of the input process.
  - The approach to modeling the network composed of M/M/c systems is to model each individual mode as if it were independent of all other nodes using as inputs to each node the same arrival process as to the first node.
- Example 5.2 Consider a problem of patients in a emergency room. We would like to know the average number of patients within the facility at any one time and the average time that a patient spends in the emergency room.



Patients — Poisson process with a mean rate of 4

A single clerk — Exponential distribution of 4 minutes per patient (M/M/1)

A triage nurse — Exponential distribution of 10 minutes per patient (M/M/1)

Two doctors — Exponential distribution of 24 minutes per patient with a doctor(M/M/2)

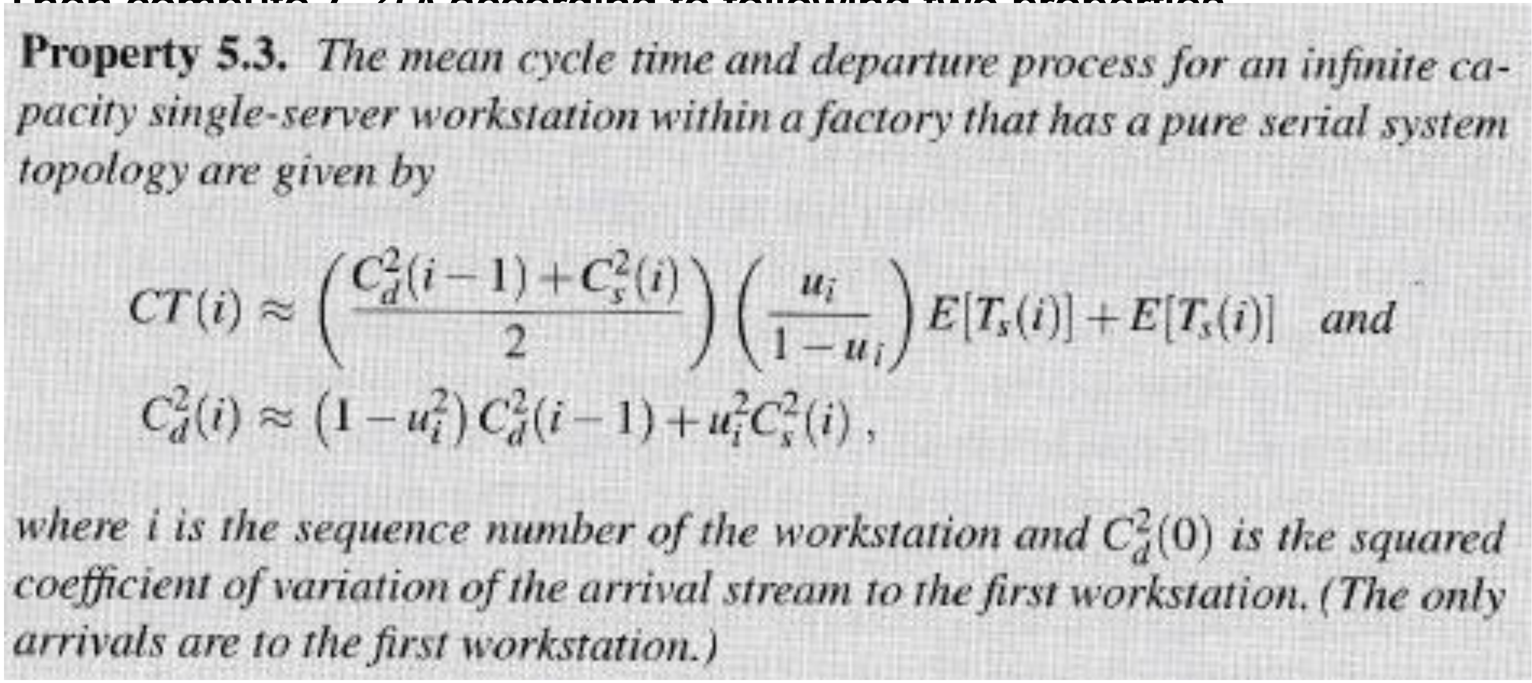
Solution:

- Arrival rate: λ=4.
- Because $E[T_a]=E[T_d]$ (according to Property 5.1), M/M/c systems have exponential inter-departure times.
- Since each of the three nodes is an infinite capacity exponential system, the system can be analyzed as three independent single node systems.
  - The first node: u_1=4/15, the average number of patients is WIP(1)=u_1/(1-u_1)=4/11
  - The second node: u_2=2/3, the $\left(\frac{C_a^2+C_s^2}{2}\right)\left(\frac{u^{\sqrt{2c+2}-1}}{c(1-u)}\right)E[T_s]$ patients is WIP(2)=u_2/(1-u_2)=2
  - The third node: u_2=4/5, $CT_q(3)$= =42.67min, $CT(3)$=1.11hr

    $WIP(3)= \lambda \times CT$=4.44
- Thus, the total number in the emergency room is

- The analysis approach for general systems is based on the concept that a system’s performance can be adequately approximated by separating the system into individual workstations.
- The performance characteristics of the individual workstations are computed separately and then these results recombined for the total system behavior.
- This decomposition approach is fundamental to the approximation of general network configurations:
    - Property 5.2 is an approximation.
    - The successive inter-departure times are not independent except for the M/M/c/∞ case.
- The parameter set required by the decomposition approach is ($E[T_s(i)]$, $C_s^2(i)$, c_i, $E[T_a(i)]$, $C_a^2(i)$) for each workstation i. The first three parameters are specified data for the workstation. The last two ones are for the job arrival stream into the workstation, which need to be estimated from the departure flows from the upstream workstations.

- The departure stream characteristics for each workstation consists of the mean inter-arrival time and the squared coefficient of variation of these times.
  - For a serial system in steady state, $E[T_a(i)]=E[T_a(1)]$ for all workstations i=1,2,…,n (the assumption of no losses, no reworks, and one external inflow point).
  - Then compute $C_d^2(i)$ according to following two properties

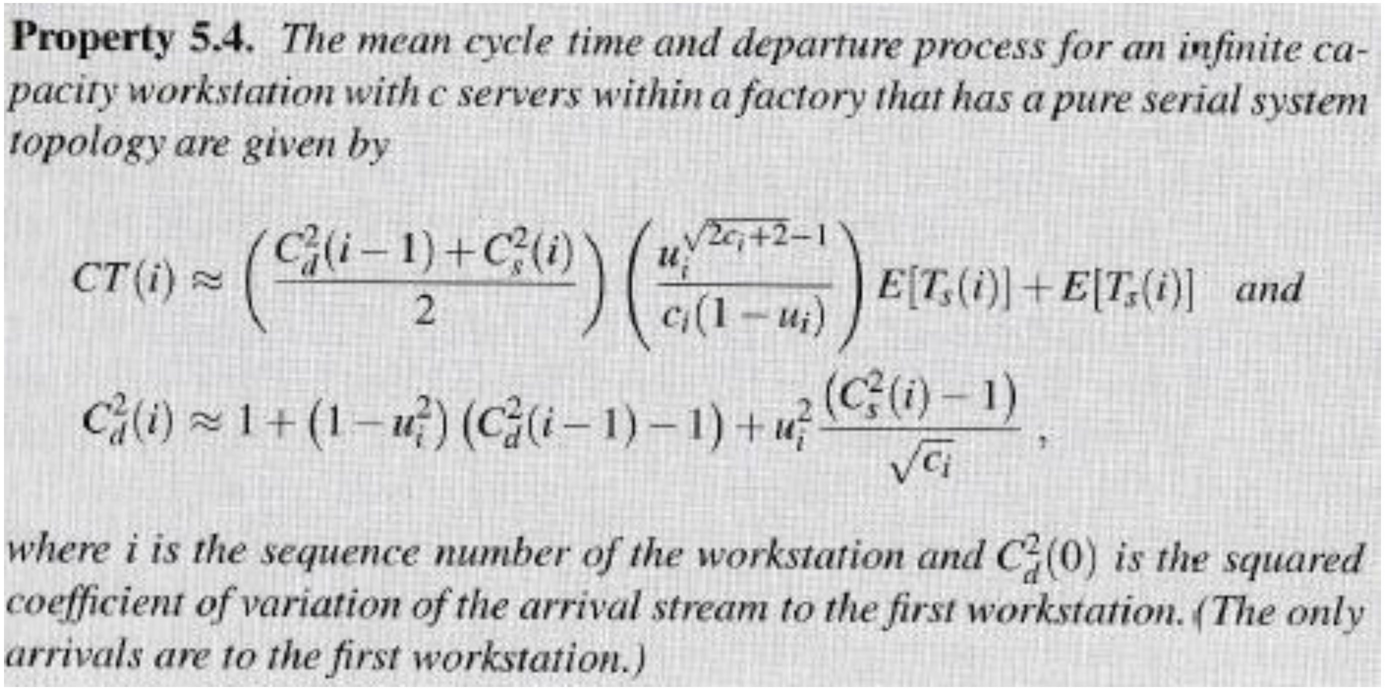


Property 5.3. *The mean cycle time and departure process for an infinite capacity single-server workstation within a factory that has a pure serial system topology are given by*

$$CT(i) \approx \left(\frac{C_d^2(i-1)+C_s^2(i)}{2}\right)\left(\frac{u_i}{1-u_i}\right)E[T_s(i)]+E[T_s(i)] \quad \textit{and}$$

$$C_d^2(i) \approx \left(1-u_i^2\right)C_d^2(i-1)+u_i^2C_s^2(i)\,,$$

where i is the sequence number of the workstation and $C_d^2(0)$ is the squared coefficient of variation of the arrival stream to the first workstation. (The only arrivals are to the first workstation.)



Property 5.4. *The mean cycle time and departure process for an infinite capacity workstation with c servers within a factory that has a pure serial system topology are given by*

$$CT(i) \approx \left(\frac{C_d^2(i-1)+C_s^2(i)}{2}\right)\left(\frac{u_i^{\sqrt{2c_i+2}-1}}{c_i(1-u_i)}\right)E[T_s(i)]+E[T_s(i)] \quad \textit{and}$$

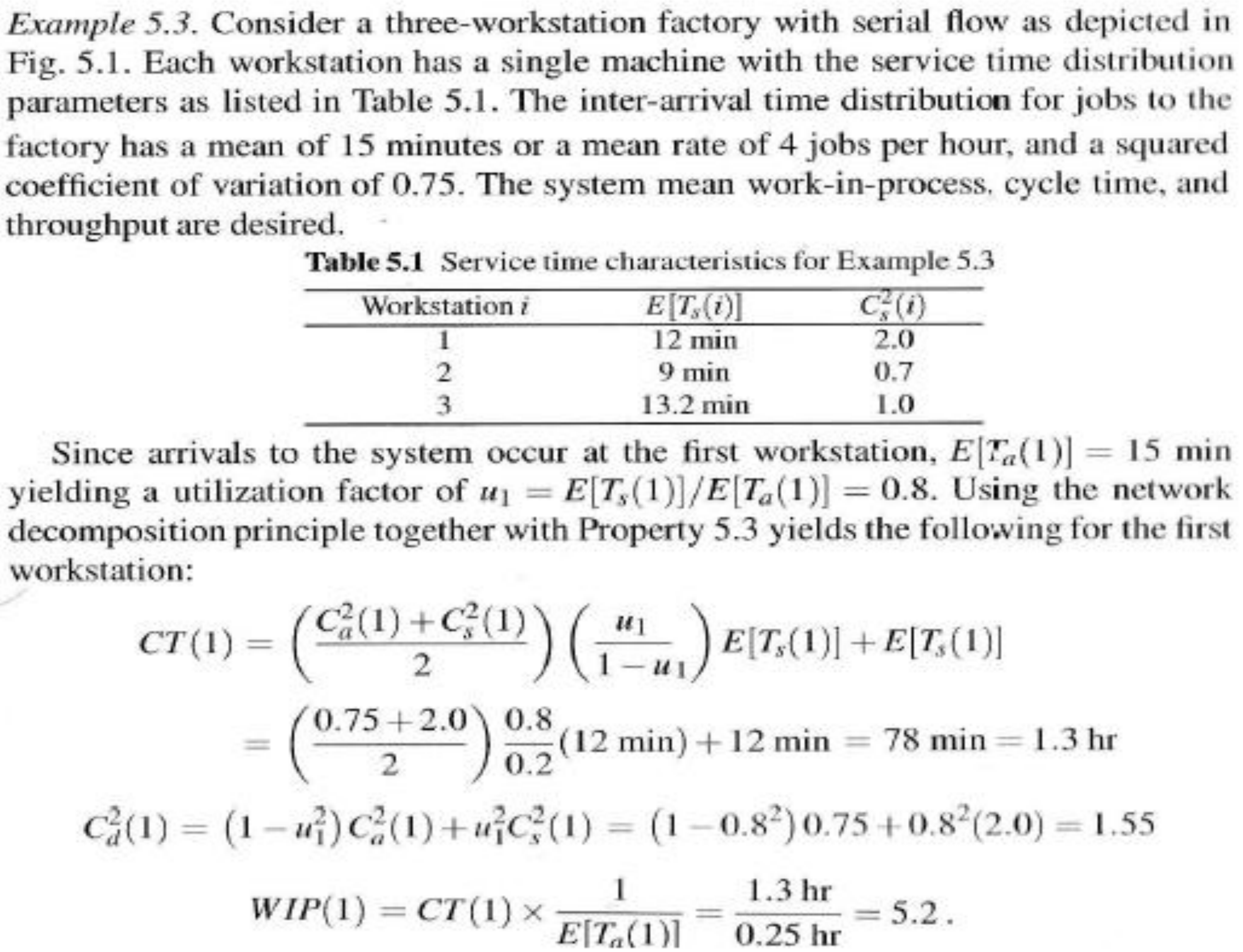
$$C_d^2(i) \approx 1+(1-u_i^2)\left(C_d^2(i-1)-1\right)+u_i^2\frac{(C_s^2(i)-1)}{\sqrt{c_i}},$$

where i is the sequence number of the workstation and $C_d^2(0)$ is the squared coefficient of variation of the arrival stream to the first workstation. (The only arrivals are to the first workstation.)

- Once the cycle times for the individual workstations have been obtained, the overall system performance measures can be determined by merely summing the individual workstation times. However, it is not a general computation scheme.
- General computation scheme:

$$WIP_s = \sum_{i=1}^{n} WIP_s(i) = \sum_{i=1}^{n} \frac{CT(i)}{E[T_a(i)]} \quad \text{and} \quad CT_s = E[T_a(1)] \times WIP_s \, .$$

The latter is assumed that all arrivals to the factory enter through the first workstation.



Example 5.3. Consider a three-workstation factory with serial flow as depicted in Fig. 5.1. Each workstation has a single machine with the service time distribution parameters as listed in Table 5.1. The inter-arrival time distribution for jobs to the factory has a mean of 15 minutes or a mean rate of 4 jobs per hour, and a squared coefficient of variation of 0.75. The system mean work-in-process, cycle time, and throughput are desired.

Table 5.1 Service time characteristics for Example 5.3

Workstation i	$E[T_s(i)]$	$C_s^2(i)$
1	12 min	2.0
2	9 min	0.7
3	13.2 min	1.0

Since arrivals to the system occur at the first workstation, $E[T_a(1)] = 15$ min yielding a utilization factor of $u_1 = E[T_s(1)]/E[T_a(1)] = 0.8$. Using the network decomposition principle together with Property 5.3 yields the following for the first workstation:

$$CT(1) = \left(\frac{C_a^2(1) + C_s^2(1)}{2}\right)\left(\frac{u_1}{1-u_1}\right)E[T_s(1)] + E[T_s(1)]$$

$$= \left(\frac{0.75 + 2.0}{2}\right)\frac{0.8}{0.2}(12 \text{ min}) + 12 \text{ min} = 78 \text{ min} = 1.3 \text{ hr}$$

$$C_d^2(1) = (1-u_1^2)\,C_a^2(1) + u_1^2 C_s^2(1) = (1-0.8^2)\,0.75 + 0.8^2(2.0) = 1.55$$

$$WIP(1) = CT(1) \times \frac{1}{E[T_a(1)]} = \frac{1.3 \text{ hr}}{0.25 \text{ hr}} = 5.2\,.$$

The last equation comes from the application of Little's Law, and since no jobs are lost, the throughput rate is $th = 1/E[T_a(1)]$. Notice that care must always be taken to make sure that the time units are consistent when applying Little's Law. Because this is a pure serial network, the arrival rate and throughput rate will be the same for each workstation; thus, the utilization factors for the other two workstations are $u_2 = E[T_s(2)]/E[T_a(1)] = 0.6$ and $u_3 = E[T_s(3)]/E[T_a(1)] = 0.88$. Applying Property 5.3 and Little's Law to the second and third workstations yield

$$
\begin{aligned}
CT(2) &= \left(\frac{1.55+0.7}{2}\right)\frac{0.6}{0.4}(0.15\text{ hr}) + 0.15\text{ hr} = 0.403\text{ hr}\\
C_d^2(2) &= (1-0.6^2)\,1.55 + 0.6^2(0.7) = 1.244\\
WIP(2) &= CT(2)/E[T_a(1)] = 1.613 \quad \text{and}\\
CT(3) &= \left(\frac{1.244+1.0}{2}\right)\frac{0.88}{0.12}(0.22\text{ hr}) + 0.22\text{ hr} = 2.030\text{ hr}\\
C_d^2(3) &= (1-0.88^2)\,1.244 + 0.88^2(1.0) = 1.055\\
WIP_s(3) &= CT(3)/E[T_a(1)] = 8.121\,.
\end{aligned}
$$

Finally, the total factory performance characteristics for this serial system are

$$WIP_s = 5.200 + 1.613 + 8.121 = 14.933\text{ jobs}$$

$$th_s = \frac{1}{E[T_a(1)]} = 4/\text{hr} \qquad CT_s = \frac{WIP_s}{th_s} = 3.733\text{ hr}\,.$$

1.5.3 Nonserial Network Models

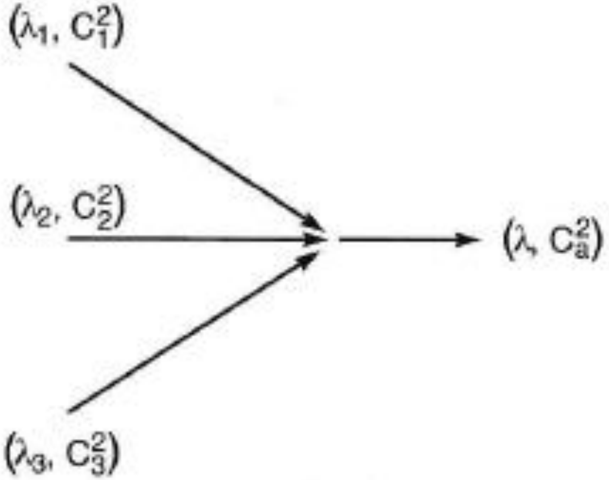
- Many production systems have more than one inflow point into the production system, such as the rework of the defective or broken jobs.
  - These rework jobs will not necessarily enter the production line at the same point as a new job.
  - If a defect is found during inspection after partially completing production, it may be sent to a rework station and then re-enter the production sequence at the appropriate point.
- To study factory structures that are more realistic than pure serial systems, two additional structures must be studied:

(1) the merging of streams entering a workstation;

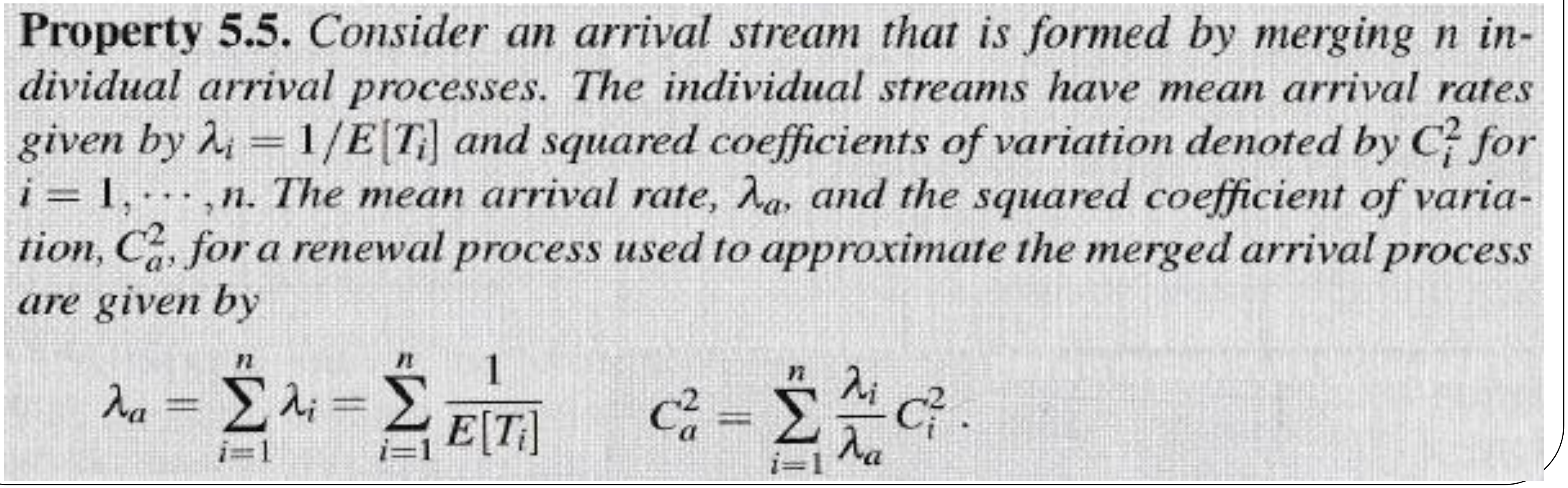
(2) the splitting of output streams that come from a single workstation but are routed to more than one workstation.

1.5.3.1 Merging Inflow Streams

- The process of merging inflow streams is technically called a superposition of the individual inter-arrival processes.

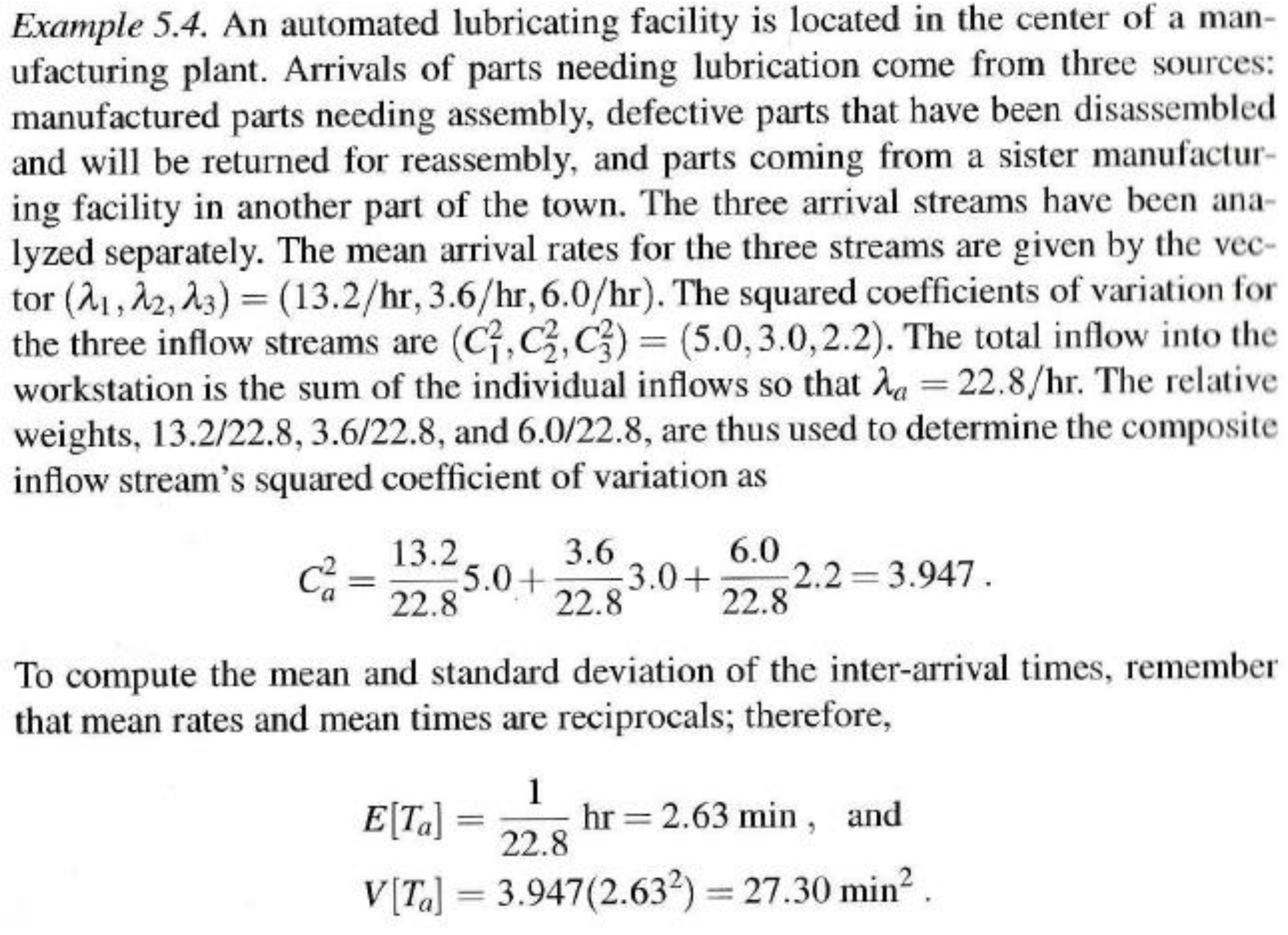


Definition 5.1. A *renewal process* is the process formed by the sum of nonnegative random variables that are independent and identically distributed. If the random variables forming the sum are exponentially distributed, the renewal process is called a *Poisson process*.



Property 5.5. *Consider an arrival stream that is formed by merging n individual arrival processes. The individual streams have mean arrival rates given by $\lambda_i = 1/E[T_i]$ and squared coefficients of variation denoted by C_i^2 for $i = 1, \cdots, n$. The mean arrival rate, λ_a, and the squared coefficient of variation, C_a^2, for a renewal process used to approximate the merged arrival process are given by*

$$\lambda_a = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \frac{1}{E[T_i]} \qquad C_a^2 = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_a} C_i^2 .$$



Example 5.4. An automated lubricating facility is located in the center of a manufacturing plant. Arrivals of parts needing lubrication come from three sources: manufactured parts needing assembly, defective parts that have been disassembled and will be returned for reassembly, and parts coming from a sister manufacturing facility in another part of the town. The three arrival streams have been analyzed separately. The mean arrival rates for the three streams are given by the vector $(\lambda_1, \lambda_2, \lambda_3) = (13.2/\text{hr}, 3.6/\text{hr}, 6.0/\text{hr})$. The squared coefficients of variation for the three inflow streams are $(C_1^2, C_2^2, C_3^2) = (5.0, 3.0, 2.2)$. The total inflow into the workstation is the sum of the individual inflows so that $\lambda_a = 22.8/\text{hr}$. The relative weights, 13.2/22.8, 3.6/22.8, and 6.0/22.8, are thus used to determine the composite inflow stream's squared coefficient of variation as

$$C_a^2 = \frac{13.2}{22.8}5.0 + \frac{3.6}{22.8}3.0 + \frac{6.0}{22.8}2.2 = 3.947\,.$$

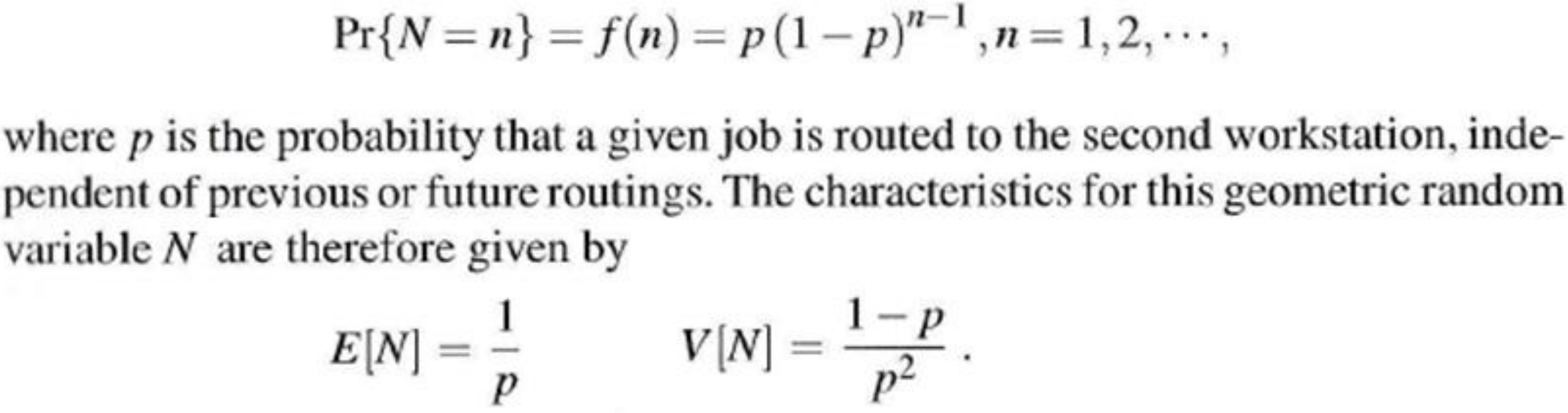
To compute the mean and standard deviation of the inter-arrival times, remember that mean rates and mean times are reciprocals; therefore,

$$E[T_a] = \frac{1}{22.8}\,\text{hr} = 2.63\,\text{min}\,, \quad \text{and}$$

$$V[T_a] = 3.947(2.63^2) = 27.30\,\text{min}^2\,.$$

1.5.3.2 Random Splitting of the Departure Stream

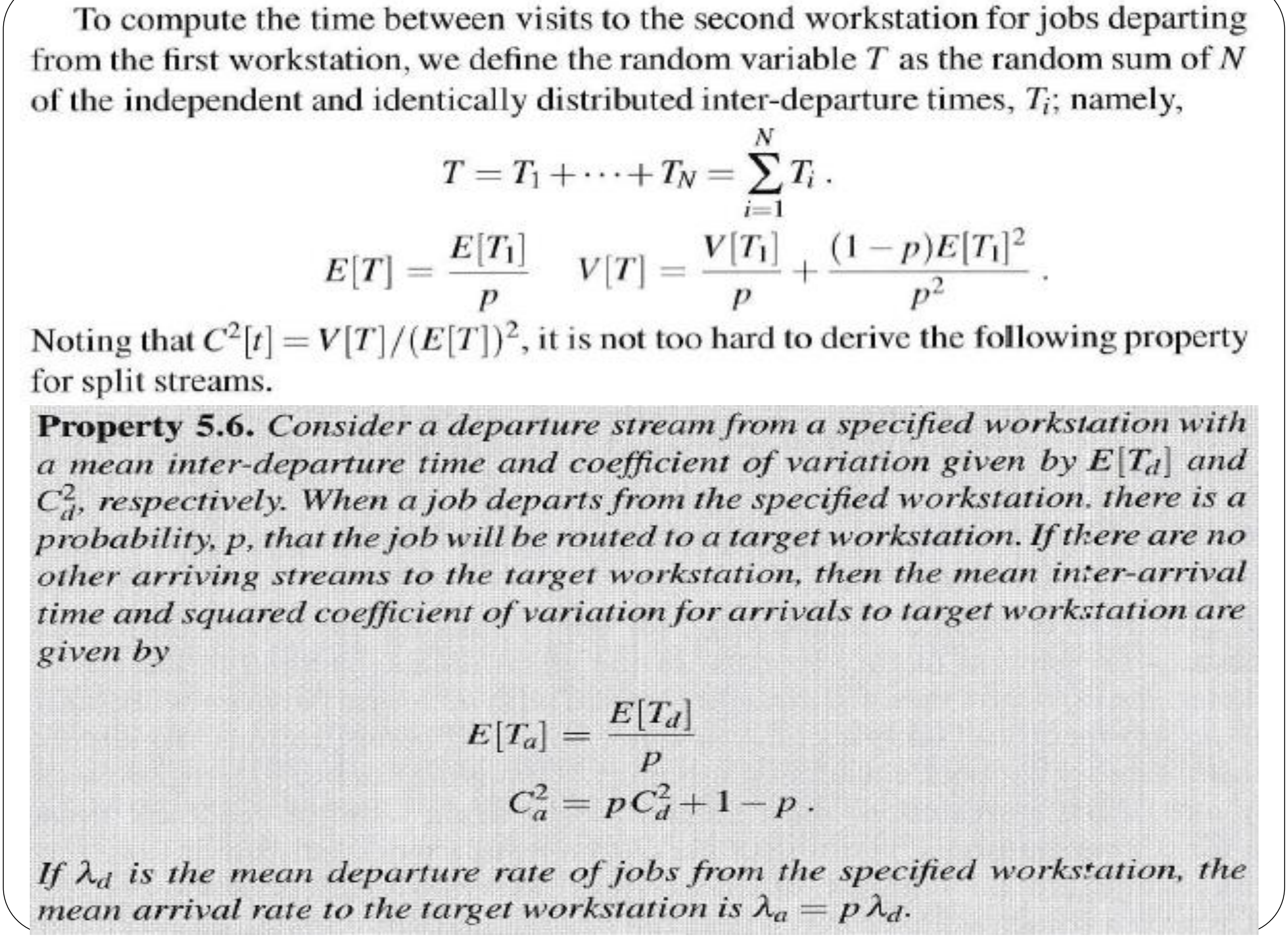
- Jobs that exit from a workstation can be transferred to different workstations based on several possibilities.
  - Multiple products can be made by specializing a partially processed product.
  - Quality control testing with good items continue on their normal route and bad ones being reworked or corrected at a different workstation before continuing normal processing.
- Assume *p* is the probability that output from one workstation is directed as an arrival process to a second workstation, and *N* is the number of departures from the first workstation between arrivals to the second workstation. Thus, the probability mass function of *N* is given by



$$\Pr\{N = n\} = f(n) = p(1-p)^{n-1}, n = 1, 2, \cdots,$$

where p is the probability that a given job is routed to the second workstation, independent of previous or future routings. The characteristics for this geometric random variable N are therefore given by

$$E[N] = \frac{1}{p} \qquad V[N] = \frac{1-p}{p^2}.$$



To compute the time between visits to the second workstation for jobs departing from the first workstation, we define the random variable T as the random sum of N of the independent and identically distributed inter-departure times, T_i; namely,

$$T = T_1 + \cdots + T_N = \sum_{i=1}^{N} T_i \, .$$

$$E[T] = \frac{E[T_1]}{p} \quad V[T] = \frac{V[T_1]}{p} + \frac{(1-p)E[T_1]^2}{p^2} \, .$$

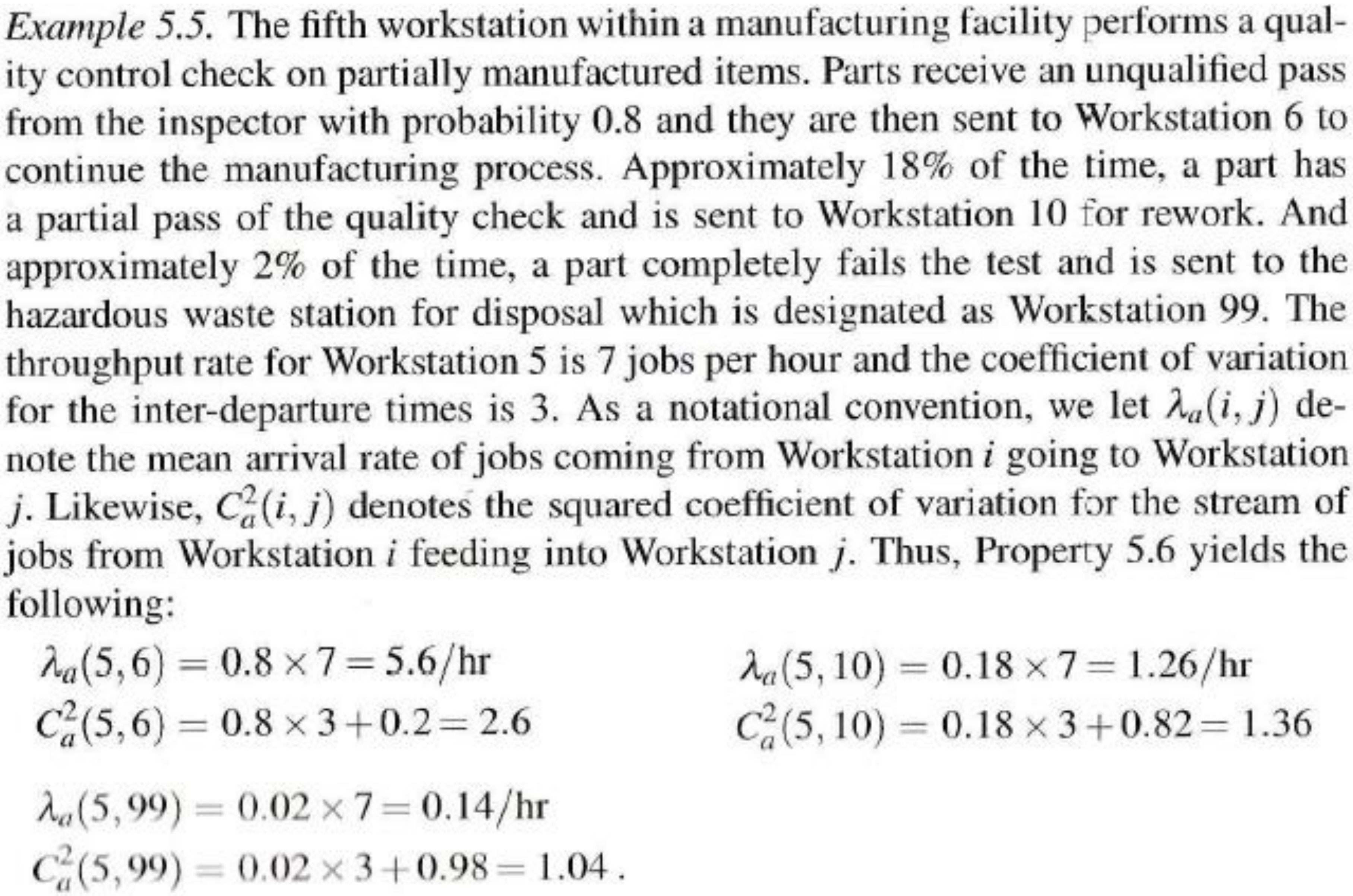
Noting that $C^2[t] = V[T]/(E[T])^2$, it is not too hard to derive the following property for split streams.

Property 5.6. *Consider a departure stream from a specified workstation with a mean inter-departure time and coefficient of variation given by $E[T_d]$ and C_d^2, respectively. When a job departs from the specified workstation, there is a probability, p, that the job will be routed to a target workstation. If there are no other arriving streams to the target workstation, then the mean inter-arrival time and squared coefficient of variation for arrivals to target workstation are given by*

$$E[T_a] = \frac{E[T_d]}{p}$$

$$C_a^2 = p\,C_d^2 + 1 - p \, .$$

If λ_d is the mean departure rate of jobs from the specified workstation, the mean arrival rate to the target workstation is $\lambda_a = p\,\lambda_d$.



Example 5.5. The fifth workstation within a manufacturing facility performs a quality control check on partially manufactured items. Parts receive an unqualified pass from the inspector with probability 0.8 and they are then sent to Workstation 6 to continue the manufacturing process. Approximately 18% of the time, a part has a partial pass of the quality check and is sent to Workstation 10 for rework. And approximately 2% of the time, a part completely fails the test and is sent to the hazardous waste station for disposal which is designated as Workstation 99. The throughput rate for Workstation 5 is 7 jobs per hour and the coefficient of variation for the inter-departure times is 3. As a notational convention, we let $\lambda_a(i,j)$ denote the mean arrival rate of jobs coming from Workstation i going to Workstation j. Likewise, $C_a^2(i,j)$ denotes the squared coefficient of variation for the stream of jobs from Workstation i feeding into Workstation j. Thus, Property 5.6 yields the following:

$$\lambda_a(5,6) = 0.8 \times 7 = 5.6/\text{hr} \qquad \lambda_a(5,10) = 0.18 \times 7 = 1.26/\text{hr}$$

$$C_a^2(5,6) = 0.8 \times 3 + 0.2 = 2.6 \qquad C_a^2(5,10) = 0.18 \times 3 + 0.82 = 1.36$$

$$\lambda_a(5,99) = 0.02 \times 7 = 0.14/\text{hr}$$

$$C_a^2(5,99) = 0.02 \times 3 + 0.98 = 1.04\,.$$

Notice that as a check, the arrival rates can be summed and they must equal the departure rate from the original stream before it was split. Such a property does not hold for the squared coefficients of variations.

1.5.4 The General Network Approximation Model

- External flows into any one of the workstations, rework branches, splitting of the output from a workstation to different next workstations, etc.

1.5.4.1 Computing Workstation Mean Arrival Rates

- Consider a simple two workstation example.

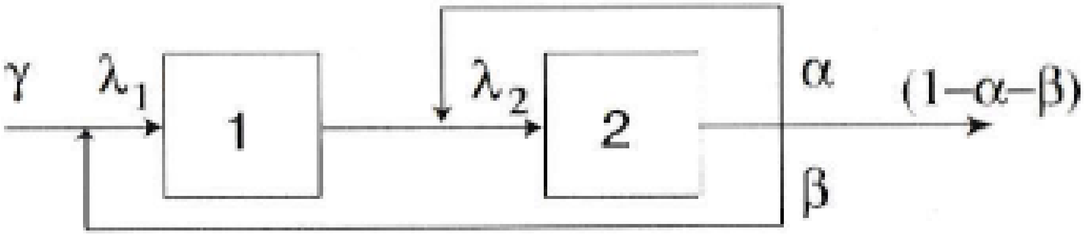


Fig. 5.3 Example of a non-serial factory model

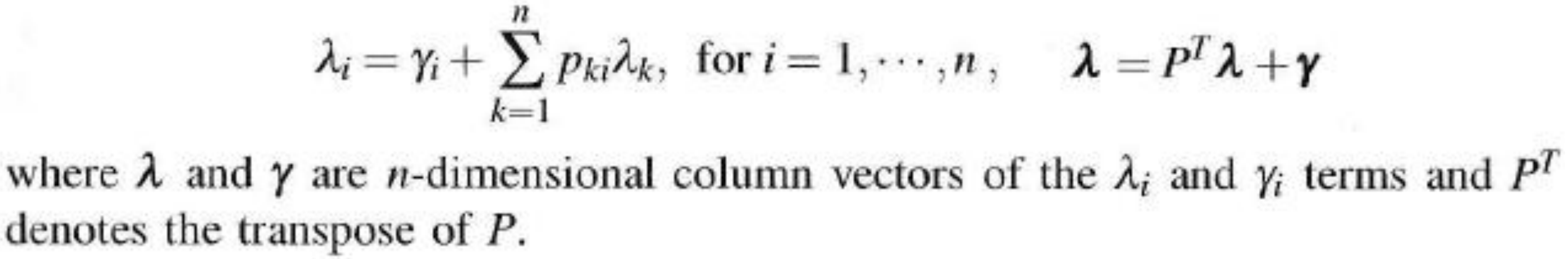
Arrivals from an external source enter the first workstation with a mean rate of *γ*.

Feedback from workstation 2 with probability *β*.

$$\begin{aligned}\lambda_1 &= \gamma+\beta\lambda_2,\\ \lambda_2 &= \lambda_1+\alpha\lambda_2,\end{aligned} \Rightarrow \begin{aligned}\lambda_1-\beta\lambda_2 &= \gamma,\\ -\lambda_1+(1-\alpha)\lambda_2 &= 0.\end{aligned} \Rightarrow \begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix}1 & -\beta\\ -1 & 1-\alpha\end{pmatrix}^{-1}\begin{pmatrix}\gamma\\ 0\end{pmatrix}$$

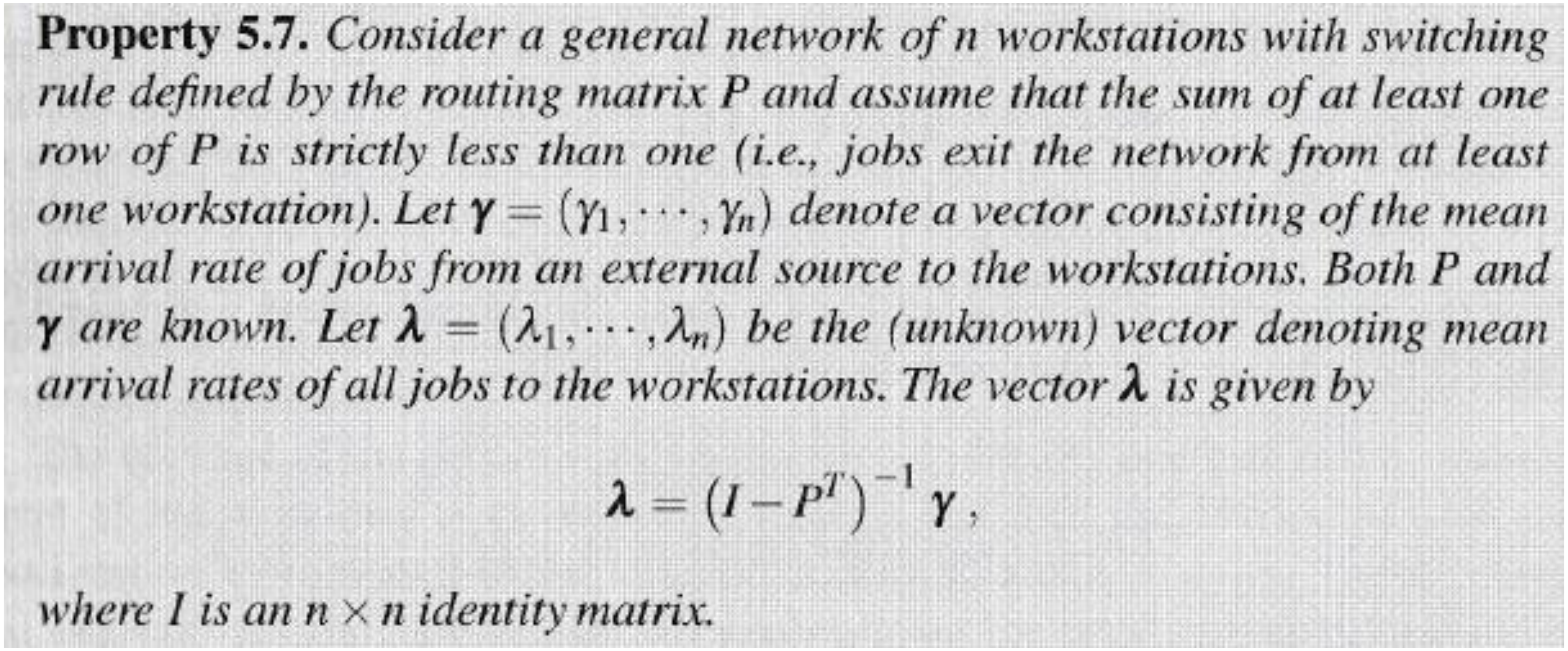
- Therefore, a system of linear equations established and solved can obtain the mean inflow rates for each workstation. To formalize for a general network application, the *switching rule* needs to be defined.

- Definition 5.2. Consider a network consisting of workstations numbered from 1 to *n*. The *switching rule* for the network is defined by an *n*×*n* matrix $P=(p_{ij})$, where $p_{i,j}$ is the probability that an arbitrary job leaving workstation *i* will be routed directly to workstation *j*. The matrix *P* is called the *routing matrix* for the network.
    - Row *i* of the *P* consists of the probabilities relating to the splitting of the outflow from workstation *i* into the various resultant successor workstation *j*.
    - Column *j* represents the probabilities that jobs leaving the various workstations go to workstation *j*.
    - Define γ_i as the external inflow rate and λ_i as the total inflow rate into workstation *i*. Therefore, the total rate into workstation I must satisfy the



$$\lambda_i = \gamma_i + \sum_{k=1}^{n} p_{ki}\lambda_k, \quad \text{for } i = 1, \cdots, n, \qquad \boldsymbol{\lambda} = P^T\boldsymbol{\lambda} + \boldsymbol{\gamma}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are *n*-dimensional column vectors of the λ_i and γ_i terms and P^T denotes the transpose of *P*.

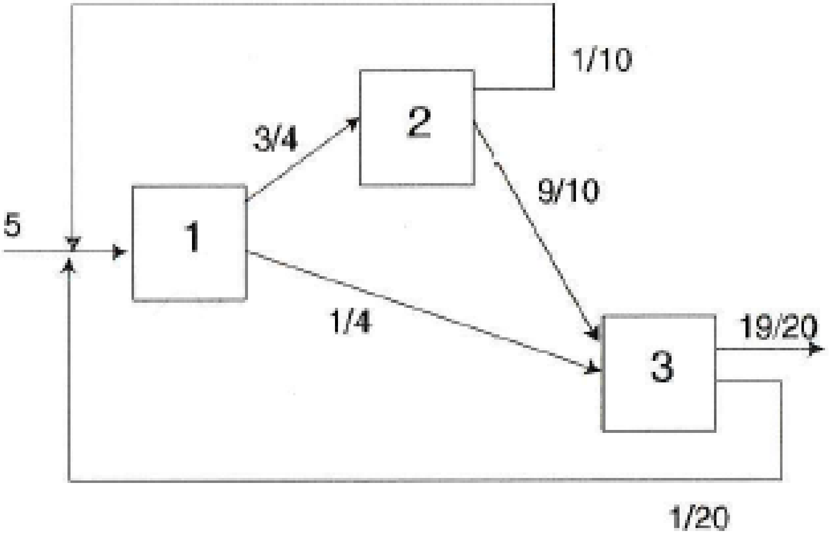


Property 5.7. *Consider a general network of n workstations with switching rule defined by the routing matrix P and assume that the sum of at least one row of P is strictly less than one (i.e., jobs exit the network from at least one workstation). Let* $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_n)$ *denote a vector consisting of the mean arrival rate of jobs from an external source to the workstations. Both P and* $\boldsymbol{\gamma}$ *are known. Let* $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_n)$ *be the (unknown) vector denoting mean arrival rates of all jobs to the workstations. The vector* $\boldsymbol{\lambda}$ *is given by*

$$\boldsymbol{\lambda} = (I - P^T)^{-1}\boldsymbol{\gamma},$$

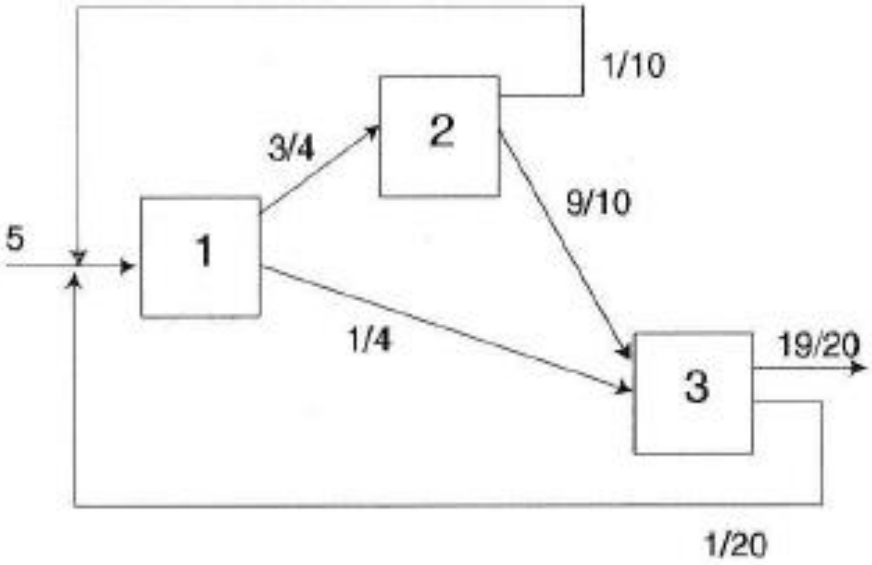
where I is an $n \times n$ *identity matrix.*

Example 5.6. Consider the factory network of workstations with the noted branching probabilities and an external flow rate into the first workstation of 5 jobs per hour.



The system of equations defining the workstation total arrival rates are

$$\lambda_1 = 5 + 0.10\lambda_2 + 0.05\lambda_3$$
$$\lambda_2 = 0 + 0.75\lambda_1$$
$$\lambda_3 = 0 + 0.25\lambda_1 + 0.90\lambda_2 .$$



This system rearranged is

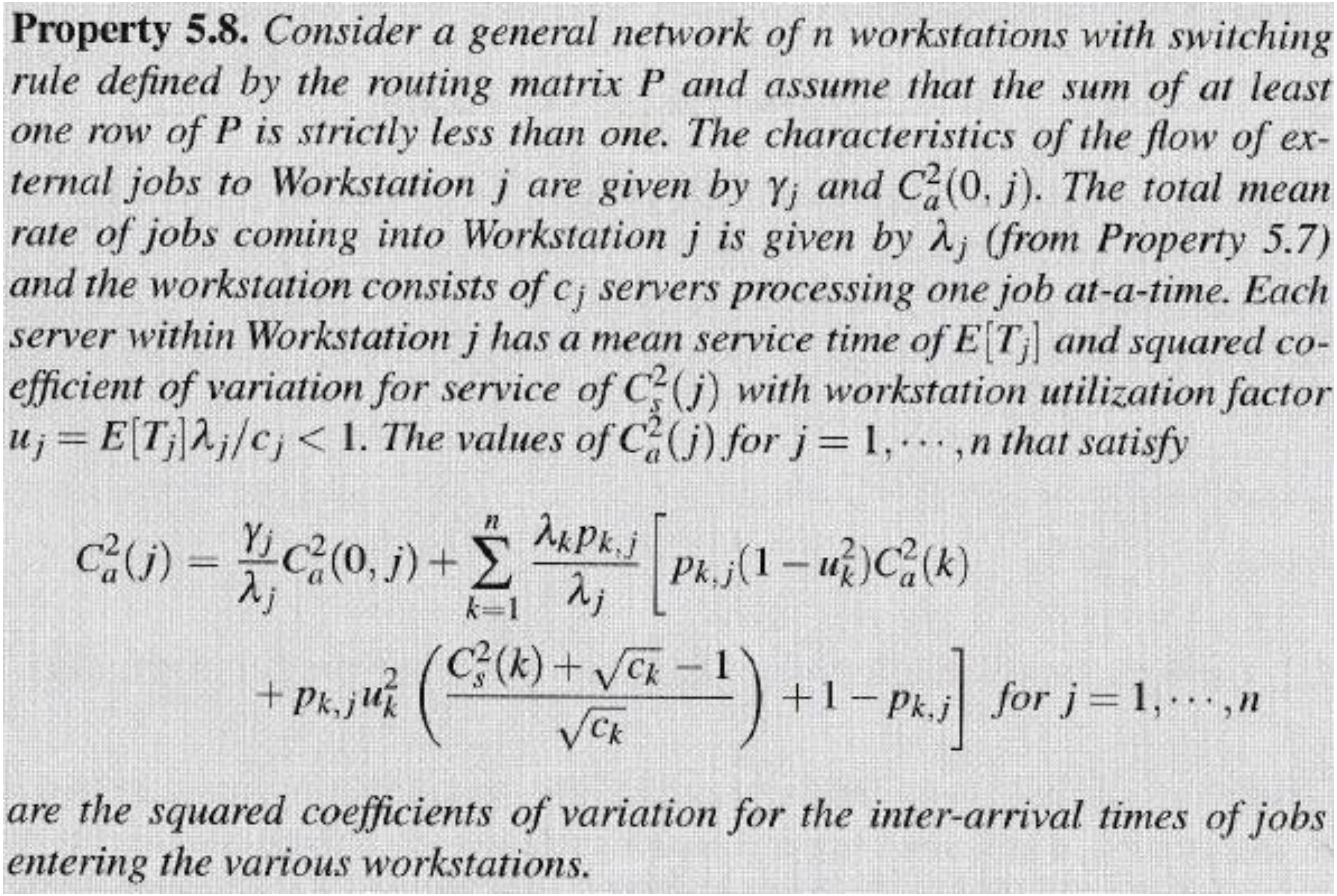
$$1\lambda_1 - 0.10\lambda_2 - 0.05\lambda_3 = 5$$
$$-0.75\lambda_1 + 1\lambda_2 + 0\lambda_3 = 0$$
$$-0.25\lambda_1 - 0.90\lambda_2 + 1\lambda_3 = 0 ,$$

which has the unique solution

$$\lambda_1 = 5.690,\ \lambda_2 = 4.267,\ \lambda_3 = 5.263 .$$

Thus, the first workstation receives 5.690 jobs per hour; 5 of these from the external source and the remaining 0.690 jobs from Workstations 2 and 3. The second workstation receives 4.267 jobs per hour, all of these from Workstation 1. The third workstation receives a total of 5.263 jobs per unit time as the combined inflow from Workstations 1 and 2.

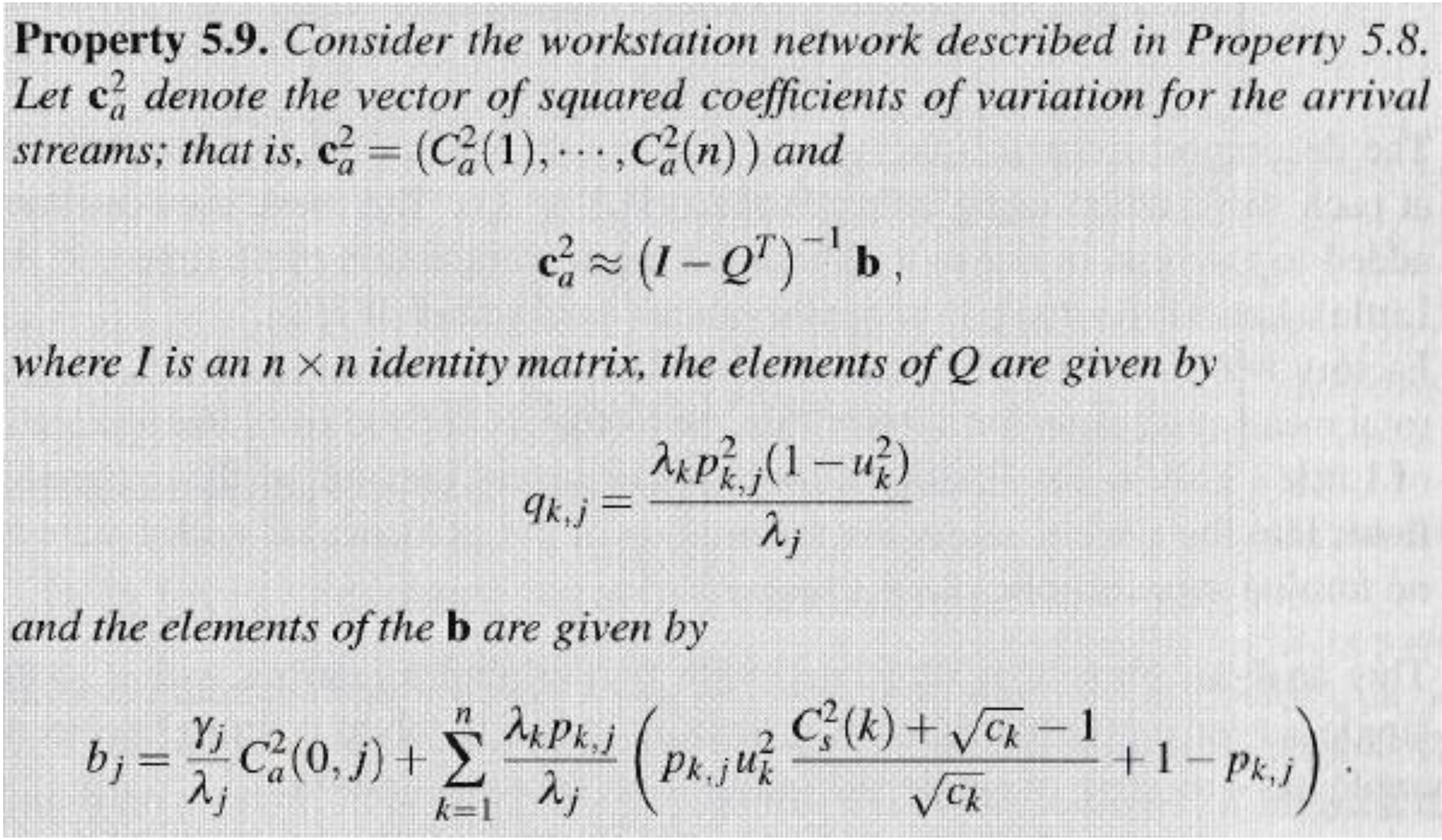
1.5.4.2 Computing Squared Coefficients of Variation for Arrivals



Property 5.8. *Consider a general network of n workstations with switching rule defined by the routing matrix P and assume that the sum of at least one row of P is strictly less than one. The characteristics of the flow of external jobs to Workstation j are given by γ_j and $C_a^2(0,j)$. The total mean rate of jobs coming into Workstation j is given by λ_j (from Property 5.7) and the workstation consists of c_j servers processing one job at-a-time. Each server within Workstation j has a mean service time of $E[T_j]$ and squared coefficient of variation for service of $C_s^2(j)$ with workstation utilization factor $u_j = E[T_j]\lambda_j/c_j < 1$. The values of $C_a^2(j)$ for $j = 1, \cdots, n$ that satisfy*

$$C_a^2(j) = \frac{\gamma_j}{\lambda_j}C_a^2(0,j) + \sum_{k=1}^{n}\frac{\lambda_k p_{k,j}}{\lambda_j}\Bigg[p_{k,j}(1-u_k^2)C_a^2(k) + p_{k,j}u_k^2\left(\frac{C_s^2(k)+\sqrt{c_k}-1}{\sqrt{c_k}}\right) + 1 - p_{k,j}\Bigg] \quad \text{for } j = 1, \cdots, n$$

are the squared coefficients of variation for the inter-arrival times of jobs entering the various workstations.



Property 5.9. *Consider the workstation network described in Property 5.8. Let* $\mathbf{c}_a^2$ *denote the vector of squared coefficients of variation for the arrival streams; that is,* $\mathbf{c}_a^2 = (C_a^2(1), \cdots, C_a^2(n))$ *and*

$$\mathbf{c}_a^2 \approx \left(I - Q^T\right)^{-1} \mathbf{b} \, ,$$

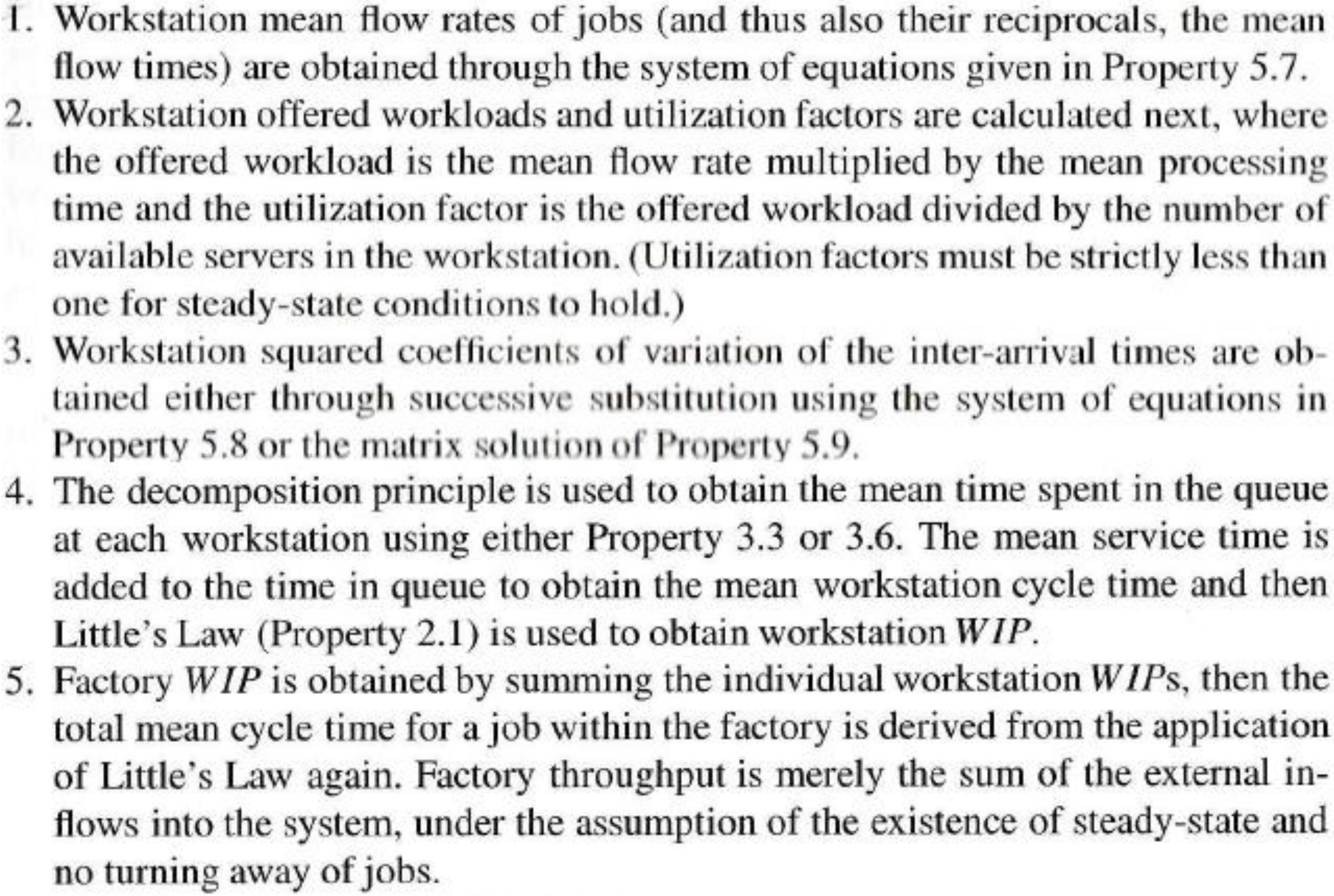
where I *is an* $n \times n$ *identity matrix, the elements of* Q *are given by*

$$q_{k,j} = \frac{\lambda_k p_{k,j}^2 (1 - u_k^2)}{\lambda_j}$$

and the elements of the **b** *are given by*

$$b_j = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \left(p_{k,j} u_k^2 \, \frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} + 1 - p_{k,j} \right) .$$

- The following is a summary of the solution procedure used to fully develop a general factory model, obtain the values of the unknown parameter sets, and derive the relevant performance measures.



1. Workstation mean flow rates of jobs (and thus also their reciprocals, the mean flow times) are obtained through the system of equations given in Property 5.7.
2. Workstation offered workloads and utilization factors are calculated next, where the offered workload is the mean flow rate multiplied by the mean processing time and the utilization factor is the offered workload divided by the number of available servers in the workstation. (Utilization factors must be strictly less than one for steady-state conditions to hold.)
3. Workstation squared coefficients of variation of the inter-arrival times are obtained either through successive substitution using the system of equations in Property 5.8 or the matrix solution of Property 5.9.
4. The decomposition principle is used to obtain the mean time spent in the queue at each workstation using either Property 3.3 or 3.6. The mean service time is added to the time in queue to obtain the mean workstation cycle time and then Little's Law (Property 2.1) is used to obtain workstation *WIP*.
5. Factory *WIP* is obtained by summing the individual workstation *WIP*s, then the total mean cycle time for a job within the factory is derived from the application of Little's Law again. Factory throughput is merely the sum of the external inflows into the system, under the assumption of the existence of steady-state and no turning away of jobs.

Step1: $\boldsymbol{\lambda} = \left(I - P^T\right)^{-1} \boldsymbol{\gamma}$

Step2: $u = E[T_s]/(c\,E[T_a])$.

Step3:

$$C_a^2(j) = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \left[p_{k,j}(1-u_k^2) C_a^2(k) + p_{k,j} u_k^2 \left(\frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} \right) + 1 - p_{k,j} \right] \quad for\ j = 1, \cdots, n$$

OR $\mathbf{c}_a^2 \approx \left(I - Q^T\right)^{-1} \mathbf{b}$

Step4:

$$CT_s(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2} \right) \left(\frac{u}{1-u} \right) E[T_s] + E[T_s]\,.$$

OR

$$CT_q(G/G/c) \approx \left(\frac{C_a^2 + C_s^2}{2} \right) \left(\frac{u^{\sqrt{2c+2}-1}}{c(1-u)} \right) E[T_s]$$

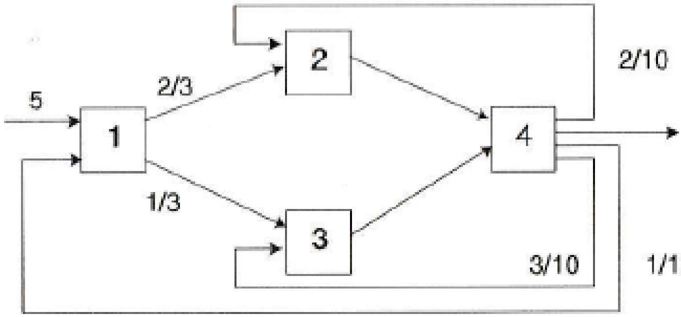
, CT_s(G/G/c)=CT_q(G/G/c)+ $E[T_s]$

Step5: $WIP = \lambda \times CT$

Example 5.7. Consider a factory that consists entirely of single-server workstations with service time data for each workstation given by Table 5.3. Arrivals from an external source enter into the factory at the first workstation, and the arrivals are exponentially distributed with a mean rate of 5 jobs per hour. After initial processing, 2/3 of the jobs are sent to Workstation 2 and 1/3 are sent to Workstation 3. After the second step of processing, jobs are tested at Workstation 4, and only 40% of the jobs are found to be acceptable. Ten percent of the completed jobs fail the testing completely and are scrapped, at which time a new job is started to replace the scrapped jobs. Fifty percent of the jobs partially fail the testing and can be reworked. Sixty percent of the partial failures are sent to Workstation 3 and the others are sent to Workstation 2. After reworking, the jobs are sent again for testing at Workstation 4 with the same percentage of passing, partially failing, and completely failing the testing. (Figure 5.5 illustrates these job flows and switching probabilities.)

Table 5.3 Workstation characteristics for Example 5.7

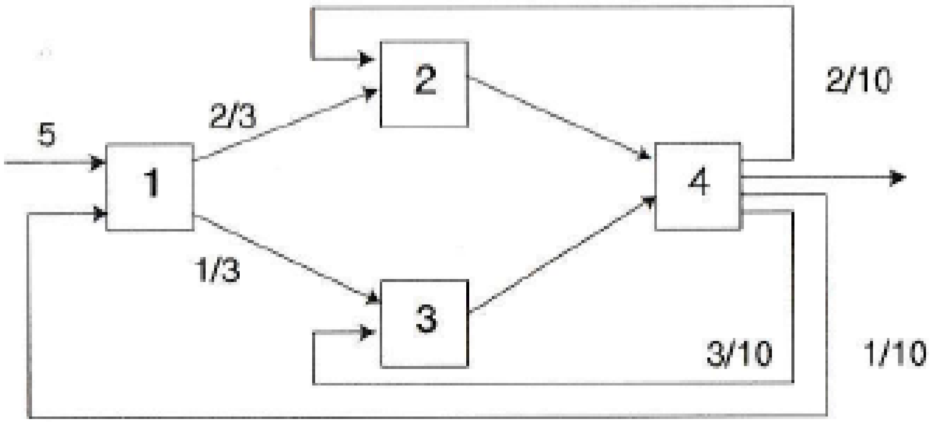
Workstation i	$E[T_s(i)]$	$C_s^2(i)$
1	7.80 min	1.0355
2	7.80 min	1.7751
3	9.60 min	0.3906
4	3.84 min	2.4414



Management is interested in the mean cycle time for jobs, factory inventory levels, and workloads at each workstation.

Step1: Workstation Arrival Rates

$$\lambda_1 = 5 + \frac{1}{10}\lambda_4$$
$$\lambda_2 = 0 + \frac{2}{3}\lambda_1 + \frac{2}{10}\lambda_4$$
$$\lambda_3 = 0 + \frac{1}{3}\lambda_1 + \frac{3}{10}\lambda_4$$
$$\lambda_4 = 0 + \lambda_2 + \lambda_3 \, .$$



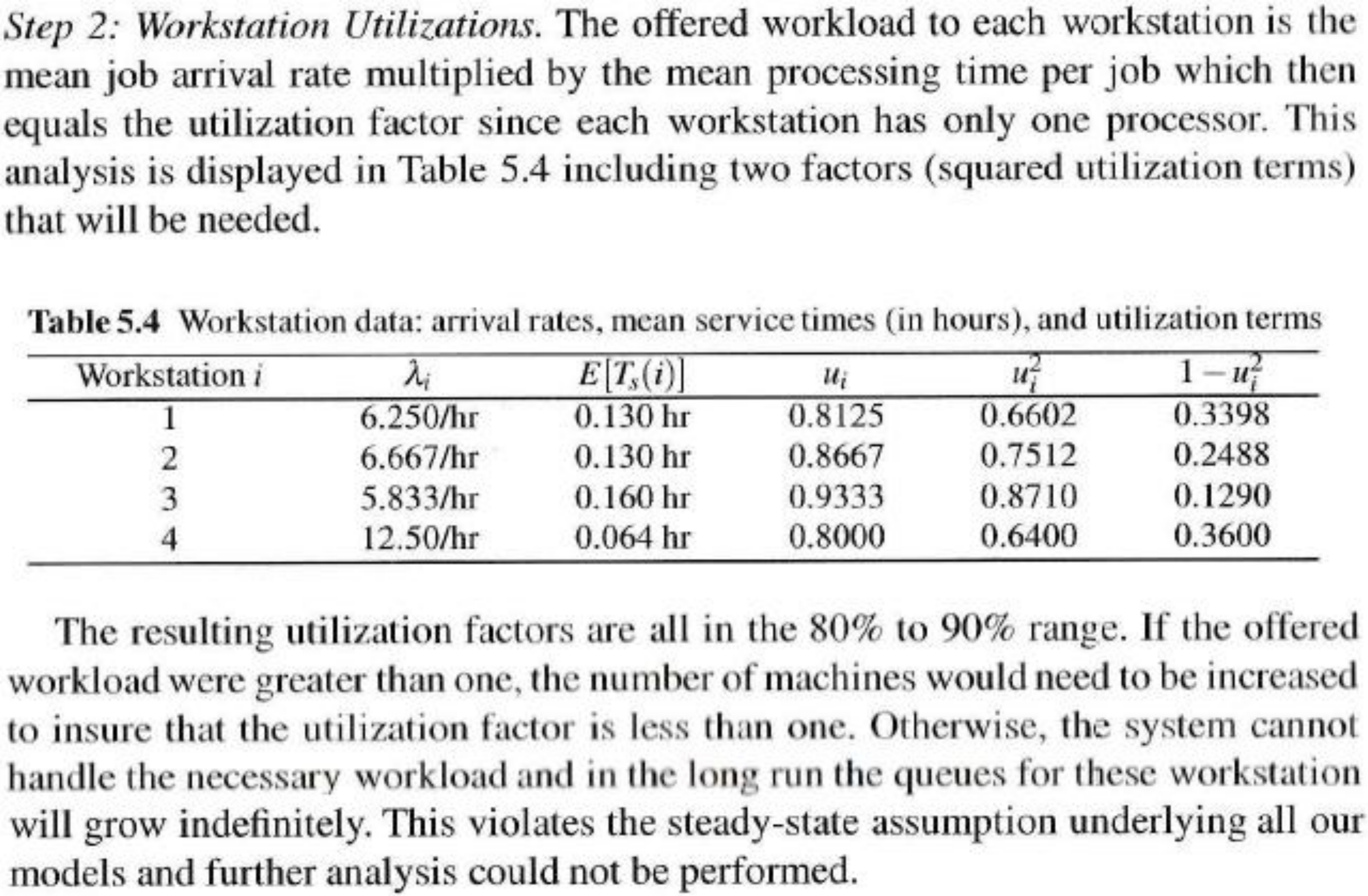
The solution to this system of equations is

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (6.25, 6.667, 5.833, 12.5) \, .$$

Thus, even though there are only 5 jobs per hour that enter into the factory, the job arrival rate into Workstation 4 is 12.5 per hour. The reason for this increase is due to the high proportion of feedback of jobs that exit Workstation 4.

$$5\left(1 + \frac{6}{10} + \left(\frac{6}{10}\right)^2 + \left(\frac{6}{10}\right)^3 + \cdots\right) = 5\left(\frac{1}{1-0.6}\right) = 12.5 \, .$$

Step2: Workstation Utilizations



Step 2: Workstation Utilizations. The offered workload to each workstation is the mean job arrival rate multiplied by the mean processing time per job which then equals the utilization factor since each workstation has only one processor. This analysis is displayed in Table 5.4 including two factors (squared utilization terms) that will be needed.

Table 5.4 Workstation data: arrival rates, mean service times (in hours), and utilization terms

Workstation i	λ_i	$E[T_s(i)]$	u_i	u_i^2	$1-u_i^2$
1	6.250/hr	0.130 hr	0.8125	0.6602	0.3398
2	6.667/hr	0.130 hr	0.8667	0.7512	0.2488
3	5.833/hr	0.160 hr	0.9333	0.8710	0.1290
4	12.50/hr	0.064 hr	0.8000	0.6400	0.3600

The resulting utilization factors are all in the 80% to 90% range. If the offered workload were greater than one, the number of machines would need to be increased to insure that the utilization factor is less than one. Otherwise, the system cannot handle the necessary workload and in the long run the queues for these workstation will grow indefinitely. This violates the steady-state assumption underlying all our models and further analysis could not be performed.

Step3: Squared Coefficients of Variation

First observe that $\gamma_2 = \gamma_3 = \gamma_4 = 0$, $\gamma_1 = 5/\text{hr}$ and $C_a^2(0,1) = 1$.

$$C_a^2(1) = \frac{5}{6.25} + \frac{12.5(0.1)}{6.25}\left[\frac{1}{10}\left(0.36\,C_a^2(4) + 0.64 \times 2.4414\right) + \frac{9}{10}\right] = 0.0072\,C_a^2(4) + 1.0112$$

$$C_a^2(2) = \frac{6.25(0.6667)}{6.6667}\left[\frac{2}{3}\left(0.3398\,C_a^2(1) + 0.6602 \times 1.0355\right) + \frac{1}{3}\right]$$
$$+ \frac{12.5(0.2)}{6.6667}\left[\frac{2}{10}\left(0.36\,C_a^2(4) + 0.64 \times 2.4414\right) + \frac{8}{10}\right] = 0.1416\,C_a^2(1) + 0.0270\,C_a^2(4) + 0.9104$$

$$C_a^2(3) = \frac{6.25(0.3333)}{5.8333}\left[\frac{1}{3}\left(0.3398\,C_a^2(1) + 0.6602 \times 1.0355\right) + \frac{2}{3}\right]$$
$$+ \frac{12.5(0.3)}{5.8333}\left[\frac{3}{10}\left(0.36\,C_a^2(4) + 0.64 \times 2.4414\right) + \frac{7}{10}\right] = 0.0405\,C_a^2(1) + 0.0694\,C_a^2(4) + 1.0708$$

$$C_a^2(4) = \frac{6.6667(1)}{12.5}\left[1\left(0.2488\,C_a^2(2) + 0.7512 \times 1.7751\right) + 0\right]$$
$$+ \frac{5.8333(1)}{12.5}\left[1\left(0.1290\,C_a^2(3) + 0.8710 \times 0.3906\right) + 0\right] = 0.1327\,C_a^2(2) + 0.0602\,C_a^2(3) + 0.8699$$

first set $\mathbf{c}_{a-step1}^2 = \left(C_a^2(1), C_a^2(2), C_a^2(3), C_a^2(4)\right)_{step1} = (1,1,1,1)$.

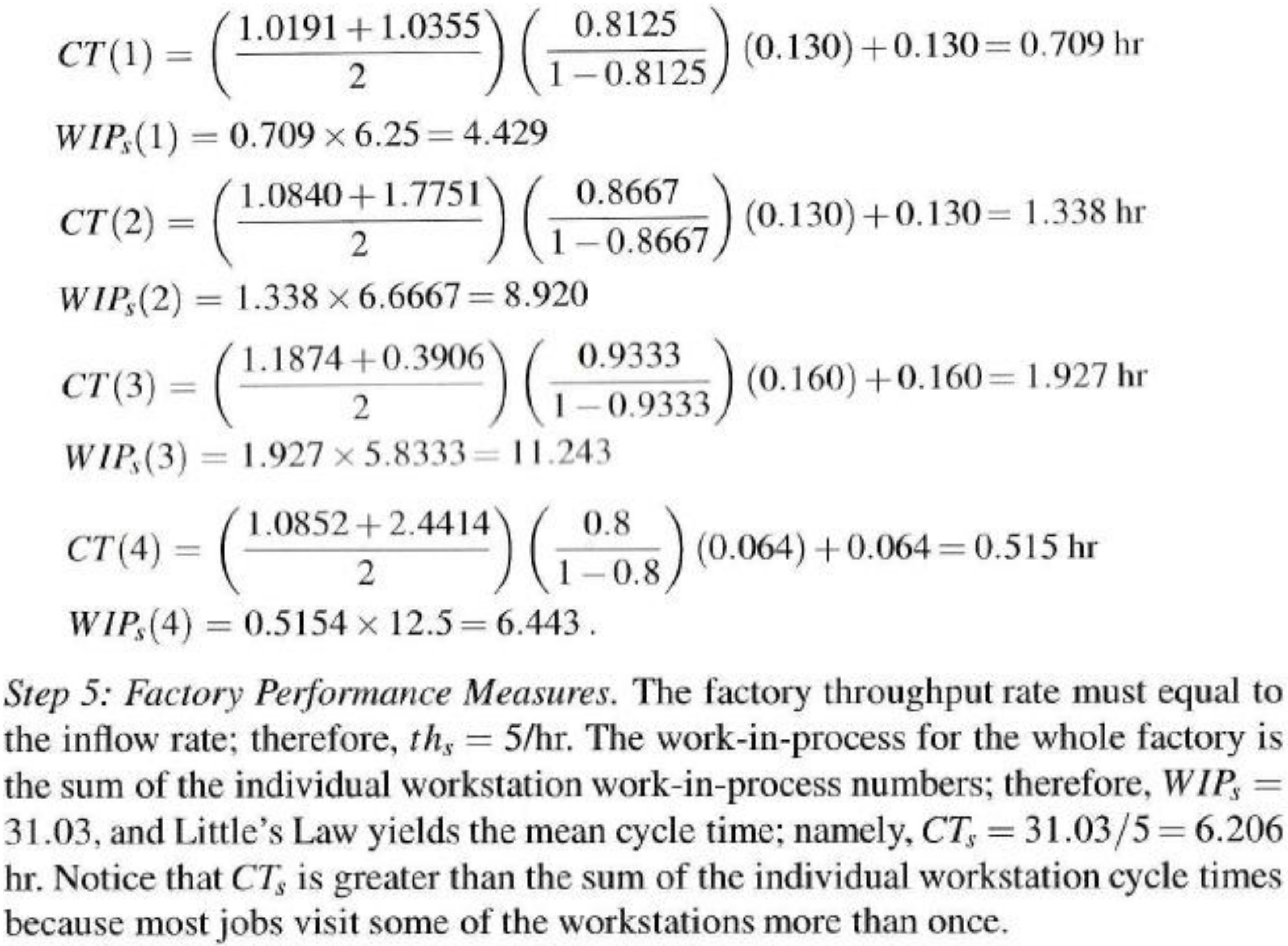
After one step of the algorithm, we have $\mathbf{c}_{a-step2}^2 = (1.0184, 1.0790, 1.1807, 1.0628)$

The next step gives $\mathbf{c}_{a-step3}^2 = (1.0189, 1.0833, 1.1858, 1.0628)$.

By the fifth iteration, the values for the squared coefficients of variation converge to

$$\mathbf{c}_{a-step5}^2 = (1.0190, 1.0840, 1.1874, 1.0852)$$

Step4: Decomposition



$$CT(1) = \left(\frac{1.0191+1.0355}{2}\right)\left(\frac{0.8125}{1-0.8125}\right)(0.130)+0.130 = 0.709 \text{ hr}$$

$$WIP_s(1) = 0.709 \times 6.25 = 4.429$$

$$CT(2) = \left(\frac{1.0840+1.7751}{2}\right)\left(\frac{0.8667}{1-0.8667}\right)(0.130)+0.130 = 1.338 \text{ hr}$$

$$WIP_s(2) = 1.338 \times 6.6667 = 8.920$$

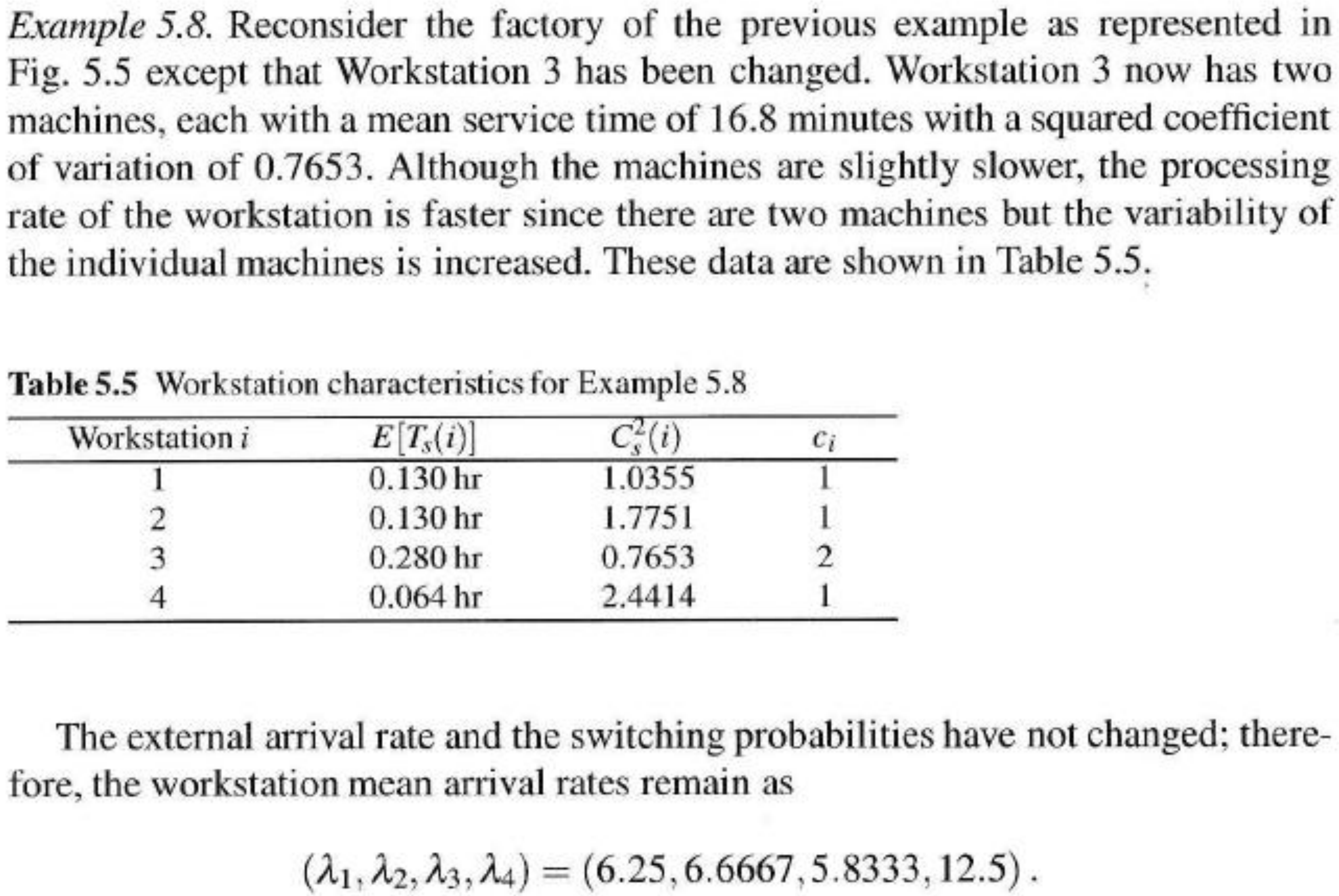
$$CT(3) = \left(\frac{1.1874+0.3906}{2}\right)\left(\frac{0.9333}{1-0.9333}\right)(0.160)+0.160 = 1.927 \text{ hr}$$

$$WIP_s(3) = 1.927 \times 5.8333 = 11.243$$

$$CT(4) = \left(\frac{1.0852+2.4414}{2}\right)\left(\frac{0.8}{1-0.8}\right)(0.064)+0.064 = 0.515 \text{ hr}$$

$$WIP_s(4) = 0.5154 \times 12.5 = 6.443\,.$$

Step 5: Factory Performance Measures. The factory throughput rate must equal to the inflow rate; therefore, $th_s = 5$/hr. The work-in-process for the whole factory is the sum of the individual workstation work-in-process numbers; therefore, $WIP_s =$ 31.03, and Little's Law yields the mean cycle time; namely, $CT_s = 31.03/5 = 6.206$ hr. Notice that CT_s is greater than the sum of the individual workstation cycle times because most jobs visit some of the workstations more than once.



Example 5.8. Reconsider the factory of the previous example as represented in Fig. 5.5 except that Workstation 3 has been changed. Workstation 3 now has two machines, each with a mean service time of 16.8 minutes with a squared coefficient of variation of 0.7653. Although the machines are slightly slower, the processing rate of the workstation is faster since there are two machines but the variability of the individual machines is increased. These data are shown in Table 5.5.

Table 5.5 Workstation characteristics for Example 5.8

Workstation i	$E[T_s(i)]$	$C_s^2(i)$	c_i
1	0.130 hr	1.0355	1
2	0.130 hr	1.7751	1
3	0.280 hr	0.7653	2
4	0.064 hr	2.4414	1

The external arrival rate and the switching probabilities have not changed; therefore, the workstation mean arrival rates remain as

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (6.25, 6.6667, 5.8333, 12.5) .$$

Since the mean arrival rates are the same in the previous example, the three unchanged workstations having the same utilization factors. Workstation 3, however, now has two servers, $c_3 = 2$, with a different mean service times so the utilization factor is recalculated as

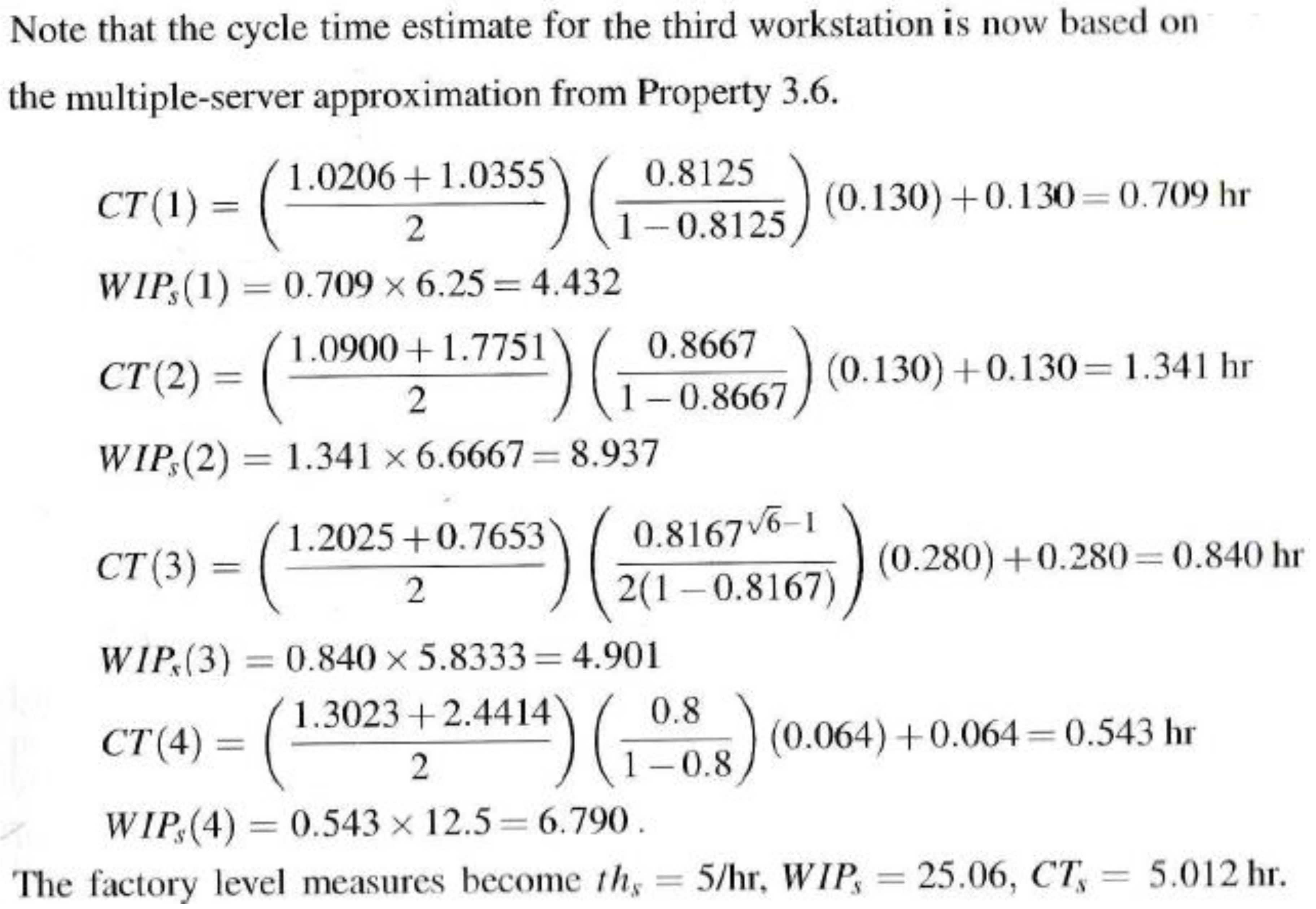
$$u_3 = \lambda_3 E[T_s(3)]/c_3 = \frac{5.8333(0.28)}{2} = 0.8167\,.$$

Since the service mechanism is changed for Workstation 3, its departure process will be changed which directly effects the arrival process for Workstation 4; therefore, the defining equation for $C_a^2(4)$ will be changed. The departure stream from Workstation 3 does not directly flow into any other workstation so all other defining equations for the squared coefficients of variation remain the same. This new equation for $C_a^2(4)$ is

$$C_a^2(4) = \frac{6.6667(1)}{12.5}\left[1\left(0.2488\,C_a^2(2) + 0.7512 \times 1.7751\right)\right] + \frac{5.8333(1)}{12.5}\left[1\left(0.3330\,C_a^2(3) + 0.6670\,\frac{0.7653 + \sqrt{2} - 1}{\sqrt{2}}\right)\right]$$

which reduces to $\quad C_a^2(4) = 0.1327\,C_a^2(2) + 0.1554\,C_a^2(3) + 0.9708\,.$

$$\mathbf{c}_a^2 = (1.0206, 1.0901, 1.2025, 1.3023)\,.$$



Note that the cycle time estimate for the third workstation is now based on the multiple-server approximation from Property 3.6.

$$CT(1) = \left(\frac{1.0206 + 1.0355}{2}\right)\left(\frac{0.8125}{1 - 0.8125}\right)(0.130) + 0.130 = 0.709 \text{ hr}$$

$$WIP_s(1) = 0.709 \times 6.25 = 4.432$$

$$CT(2) = \left(\frac{1.0900 + 1.7751}{2}\right)\left(\frac{0.8667}{1 - 0.8667}\right)(0.130) + 0.130 = 1.341 \text{ hr}$$

$$WIP_s(2) = 1.341 \times 6.6667 = 8.937$$

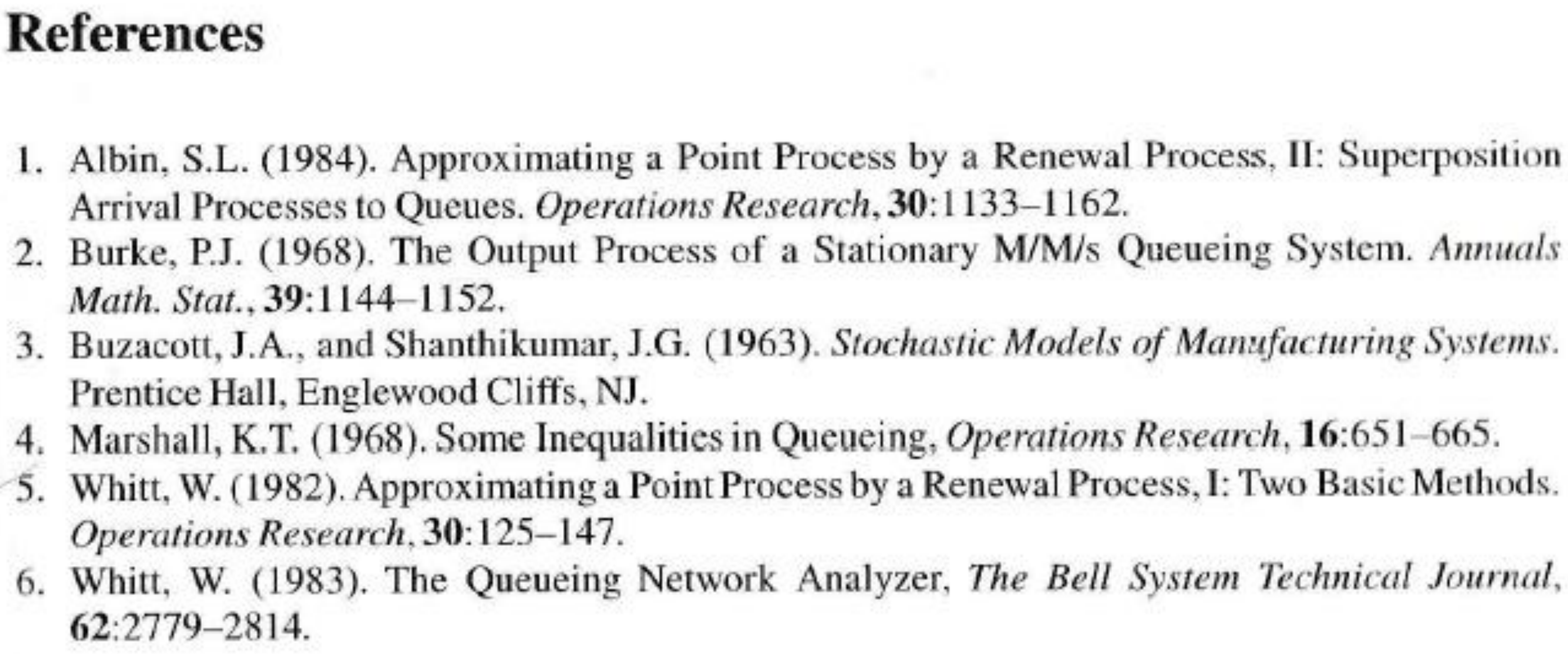
$$CT(3) = \left(\frac{1.2025 + 0.7653}{2}\right)\left(\frac{0.8167^{\sqrt{6}-1}}{2(1 - 0.8167)}\right)(0.280) + 0.280 = 0.840 \text{ hr}$$

$$WIP_s(3) = 0.840 \times 5.8333 = 4.901$$

$$CT(4) = \left(\frac{1.3023 + 2.4414}{2}\right)\left(\frac{0.8}{1 - 0.8}\right)(0.064) + 0.064 = 0.543 \text{ hr}$$

$$WIP_s(4) = 0.543 \times 12.5 = 6.790\,.$$

The factory level measures become $th_s = 5/\text{hr}$, $WIP_s = 25.06$, $CT_s = 5.012$ hr.



References

1. Albin, S.L. (1984). Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues. *Operations Research*, **30**:1133–1162.
2. Burke, P.J. (1968). The Output Process of a Stationary M/M/s Queueing System. *Annuals Math. Stat.*, **39**:1144–1152.
3. Buzacott, J.A., and Shanthikumar, J.G. (1963). *Stochastic Models of Manufacturing Systems*. Prentice Hall, Englewood Cliffs, NJ.
4. Marshall, K.T. (1968). Some Inequalities in Queueing, *Operations Research*, **16**:651–665.
5. Whitt, W. (1982). Approximating a Point Process by a Renewal Process, I: Two Basic Methods. *Operations Research*, **30**:125–147.
6. Whitt, W. (1983). The Queueing Network Analyzer, *The Bell System Technical Journal*, **62**:2779–2814.

1.6 Multi Product Factory Models

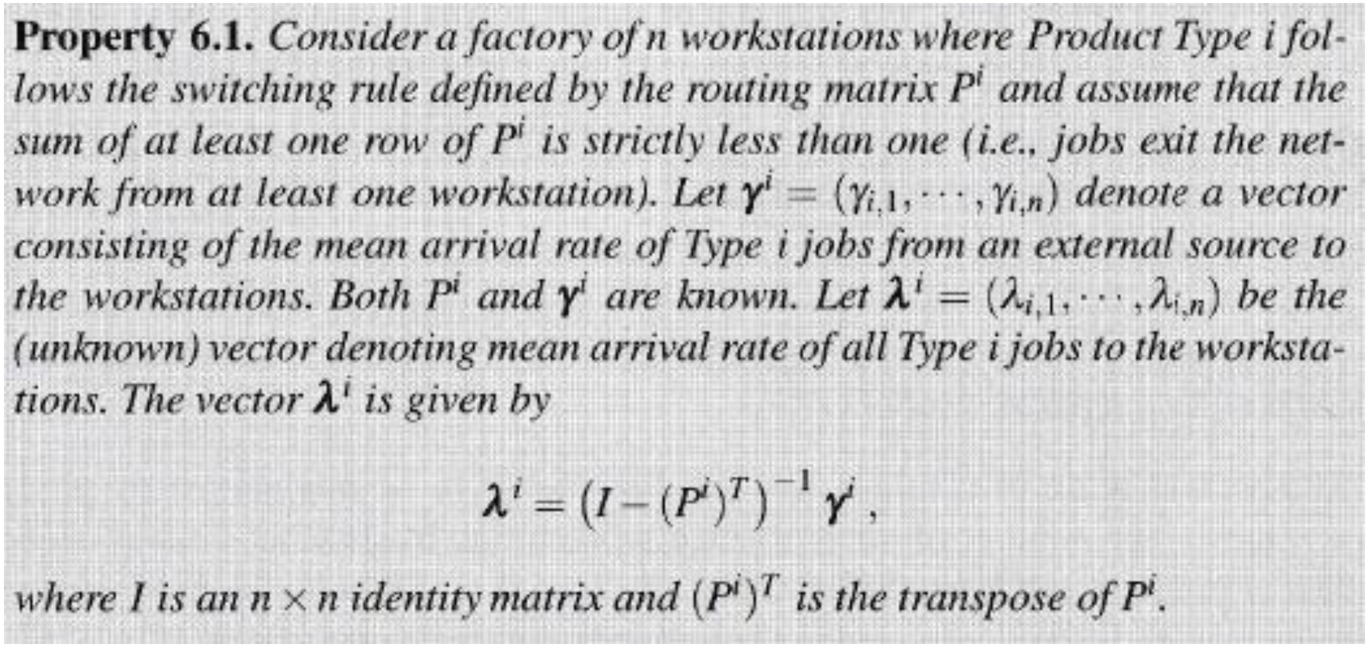
- There are two basic principles to model multiple product facilities.

(1) The workload on a workstation is the sum of all the visits multiplied by the processing time per visit.

- Definition 6.1. The *offered workload* (or simply the *workload*) of a workstation is the total amount of work that is required of a workstation per unit of time. The workload is determined by the sum of the total arrival rate (per hour) for each product type multiplied by its associated mean processing time (in hours). For purposes of determining workload, when a specific product type revisits a workstation, it is considered as a separate product type.

(2) The job flow needs to be maintained by product type. That is, the number of visits to each workstation by product class is needed. Different products can have different probabilistic flows through the production facility as well as different processing time characteristics.

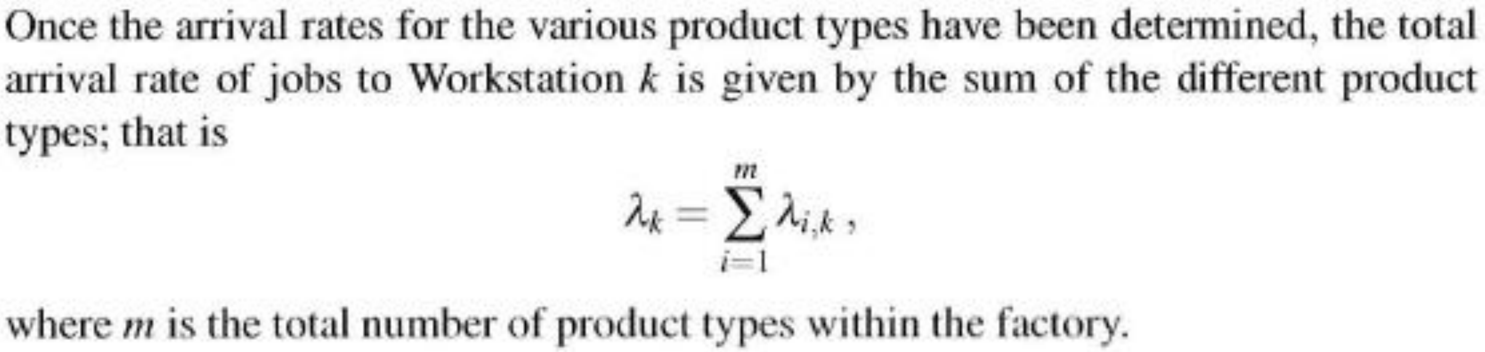
1.6.1 Product Flow Rates



Property 6.1. *Consider a factory of n workstations where Product Type i follows the switching rule defined by the routing matrix P^i and assume that the sum of at least one row of P^i is strictly less than one (i.e., jobs exit the network from at least one workstation). Let $\boldsymbol{\gamma}^i = (\gamma_{i,1}, \cdots, \gamma_{i,n})$ denote a vector consisting of the mean arrival rate of Type i jobs from an external source to the workstations. Both P^i and $\boldsymbol{\gamma}^i$ are known. Let $\boldsymbol{\lambda}^i = (\lambda_{i,1}, \cdots, \lambda_{i,n})$ be the (unknown) vector denoting mean arrival rate of all Type i jobs to the workstations. The vector $\boldsymbol{\lambda}^i$ is given by*

$$\boldsymbol{\lambda}^i = \left(I - (P^i)^T\right)^{-1} \boldsymbol{\gamma}^i ,$$

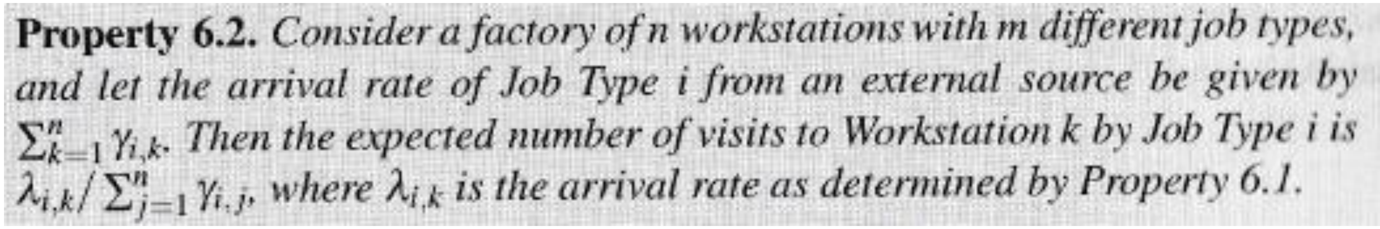
where I is an $n \times n$ identity matrix and $(P^i)^T$ is the transpose of P^i.



Once the arrival rates for the various product types have been determined, the total arrival rate of jobs to Workstation k is given by the sum of the different product types; that is

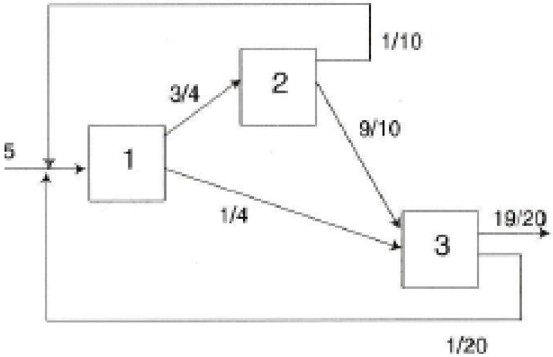
$$\lambda_k = \sum_{i=1}^{m} \lambda_{i,k} ,$$

where m is the total number of product types within the factory.



Property 6.2. *Consider a factory of n workstations with m different job types, and let the arrival rate of Job Type i from an external source be given by $\sum_{k=1}^{n} \gamma_{i,k}$. Then the expected number of visits to Workstation k by Job Type i is $\lambda_{i,k} / \sum_{j=1}^{n} \gamma_{i,j}$, where $\lambda_{i,k}$ is the arrival rate as determined by Property 6.1.*

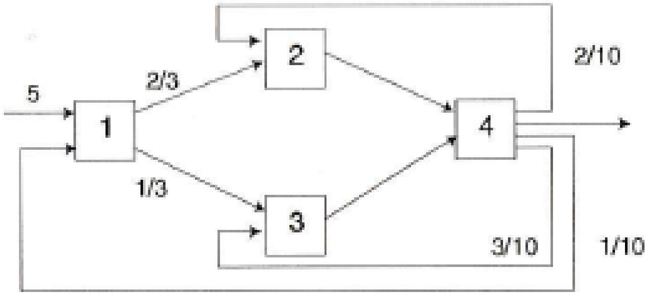
- *Example 6.1* Consider a four workstation facility that processes two products with each product arriving to the first workstation according to individual Poisson arrival streams, each at a rate of 5 per hour.



Product 1 routing structure

$$\lambda_1 = 5 + 0.10\lambda_2 + 0.05\lambda_3$$
$$\lambda_2 = 0 + 0.75\lambda_1$$
$$\lambda_3 = 0 + 0.25\lambda_1 + 0.90\lambda_2 .$$

⇒ $\boldsymbol{\lambda}^1 = (5.690, 4.267, 5.263, 0) .$



Product 2 routing structure

$$\lambda_1 = 5 + \frac{1}{10}\lambda_4$$
$$\lambda_2 = 0 + \frac{2}{3}\lambda_1 + \frac{2}{10}\lambda_4$$
$$\lambda_3 = 0 + \frac{1}{3}\lambda_1 + \frac{3}{10}\lambda_4$$
$$\lambda_4 = 0 + \lambda_2 + \lambda_3 .$$

⇒ $\boldsymbol{\lambda}^2 = (6.25, 6.667, 5.833, 12.5) .$

⇓

$\boldsymbol{\lambda} = (11.940, 10.934, 11.096, 12.5) .$

- The average number of visits of Job Type 1 to Workstation 1 is 1.138, but that to the Workstation 2 is 0.8534. The most visited workstation by a single product type is the fourth workstation that has each job type 2 visiting it an average of 2.5 times.

1.6.2 Workstation Workloads

- By Definition 6.1, the workload at Workstation *k*, WL_k, is computed as the sum of the product visits multiplied by their respectively mean processing times; that is,

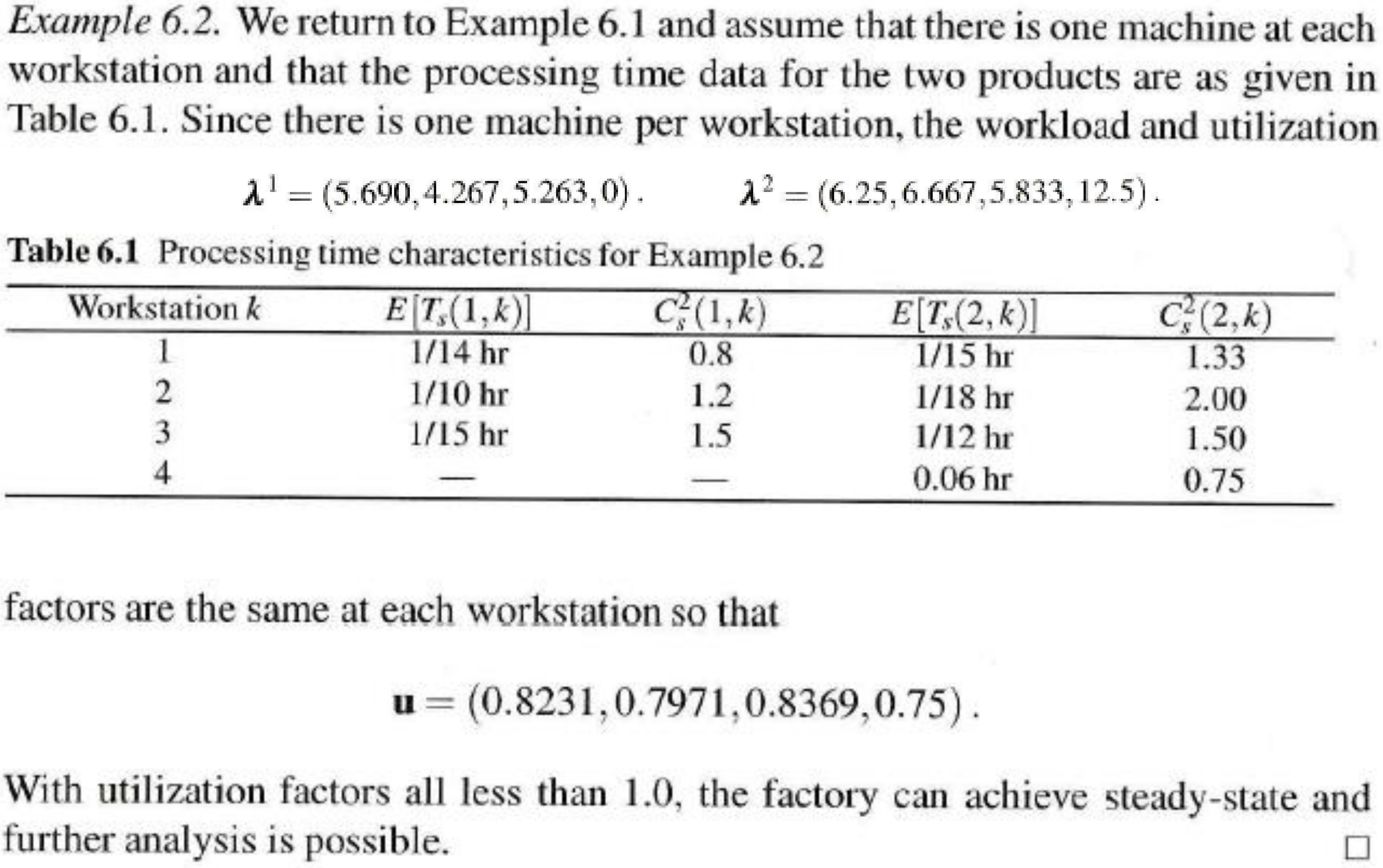
$$WL_k = \sum_{i=1}^{m} \lambda_{ik} E[T_s(i,k)]$$

where *m* is the total number of product types within the factory.

- The utilization factor, u_k, for Workstation *k* is then the workload divided by the available capacity; thus,

$$u_k = \frac{WL_k}{c_k} = \frac{\sum_{i=1}^{m} \lambda_{i,k} E[T_s(i,k)]}{c_k},$$

where c_k is the number of identical processors available at Workstation *k* to handle the workload.



Example 6.2. We return to Example 6.1 and assume that there is one machine at each workstation and that the processing time data for the two products are as given in Table 6.1. Since there is one machine per workstation, the workload and utilization

$$\boldsymbol{\lambda}^1 = (5.690, 4.267, 5.263, 0)\,. \qquad \boldsymbol{\lambda}^2 = (6.25, 6.667, 5.833, 12.5)\,.$$

Table 6.1 Processing time characteristics for Example 6.2

Workstation k	$E[T_s(1,k)]$	$C_s^2(1,k)$	$E[T_s(2,k)]$	$C_s^2(2,k)$
1	1/14 hr	0.8	1/15 hr	1.33
2	1/10 hr	1.2	1/18 hr	2.00
3	1/15 hr	1.5	1/12 hr	1.50
4	—	—	0.06 hr	0.75

factors are the same at each workstation so that

$$\mathbf{u} = (0.8231, 0.7971, 0.8369, 0.75)\,.$$

With utilization factors all less than 1.0, the factory can achieve steady-state and further analysis is possible. □

1.6.3 Service Time Characteristics

- For Workstation *k*, the service time will be the random variable $T_s(i,k)$ whenever Product *i* is being processed. The service time for an arbitrary job, independent of the job type, is the random variable denoted by $T_s(k)$. In the long-run, the probability that a given machine at Workstation *k* will be processing a Type *i* job is $\lambda_{i,k}/\lambda_k$; thus, $T_s(k)$ is a mixture of random variables sinc

$$T_s(k) = \begin{cases} T_s(1,k) & \text{with probability } \frac{\lambda_{1,k}}{\lambda_k} \\ \vdots & \\ T_s(m,k) & \text{with probability } \frac{\lambda_{m,k}}{\lambda_k} \end{cases}$$

where *m* is the total number of product types within the factory.

$$E[T_s(k)] = \sum_{i=1}^{m} \frac{\lambda_{i,k}}{\lambda_k} E[T_s(i,k)] = \frac{WL_k}{\lambda_k}$$

$V[T]=E[T^2]-E[T]^2$
$C^2[T]=V[T]/E[T]^2$
$E[T^2]=(1+C^2[T])\,E[T]^2$
$C^2[T]=E[T^2]/E[T]^2-1$

$$C_s^2(k) = \frac{\sum_{i=1}^{m} (\lambda_{i,k}/\lambda_k) E[T_s(i,k)]^2 (1 + C_s^2(i,k))}{\left(\sum_{i=1}^{m} (\lambda_{i,k}/\lambda_k) E[T_s(i,k)]\right)^2} - 1$$

Example 6.3. We are now ready to derive the mean and squared coefficients of variation for the four workstation service times using the arrival rate data of Example 6.1 and the service time data of Example 6.2.

The total arrival rate for the first workstation is 11.94/hr and thus,

$$E[T_s(1)] = \left(\frac{5.690}{11.94}\right)\frac{1}{14} + \left(\frac{6.250}{11.94}\right)\frac{1}{15} = 0.0689 \text{ hr} .$$

The computations for the squared coefficient of variation are

$$C_s^2(1) = \frac{\left(\frac{5.690}{11.94}\right)\left(\frac{1}{14}\right)^2(1+0.8) + \left(\frac{6.250}{11.94}\right)\left(\frac{1}{15}\right)^2(1+1.33)}{(0.0689)^2} - 1 = 1.0616 .$$

Note that some of the numbers used in the above equation were taken from Table 6.1. The final results for the service time characteristics for the four workstations are contained in Table 6.2. $\boldsymbol{\lambda}^1 = (5.690, 4.267, 5.263, 0)$. $\boldsymbol{\lambda}^2 = (6.25, 6.667, 5.833, 12.5)$.

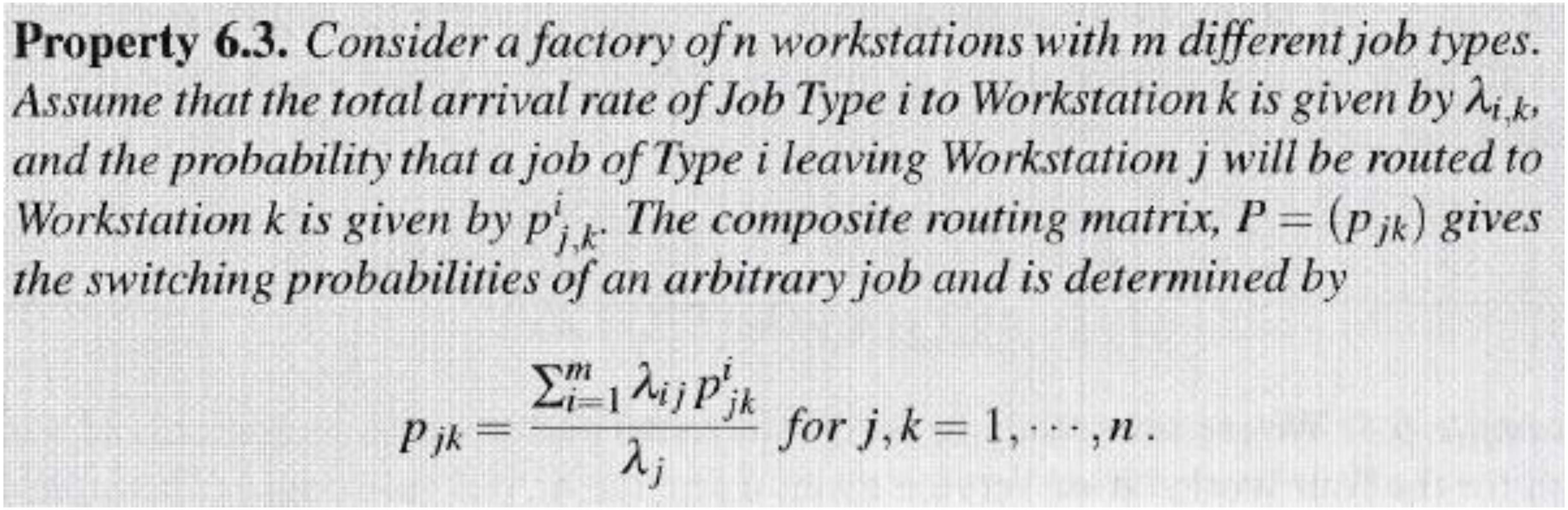
$\boldsymbol{\lambda} = (11.940, 10.934, 11.096, 12.5)$.

Table 6.2 Service time characteristics for Example 6.3

Workstation k	$E[T_s(k)]$	$C_s^2(k)$
1	0.069	1.062
2	0.073	1.678
3	0.075	1.530
4	0.060	0.750

Workstation k	$E[T_s(1,k)]$	$C_s^2(1,k)$	$E[T_s(2,k)]$	$C_s^2(2,k)$
1	1/14 hr	0.8	1/15 hr	1.33
2	1/10 hr	1.2	1/18 hr	2.00
3	1/15 hr	1.5	1/12 hr	1.50
4	—	—	0.06 hr	0.75

1.6.4 Workstation Performance Measures



Property 6.3. *Consider a factory of n workstations with m different job types. Assume that the total arrival rate of Job Type i to Workstation k is given by $\lambda_{i,k}$, and the probability that a job of Type i leaving Workstation j will be routed to Workstation k is given by $p^{i}_{j,k}$. The composite routing matrix, $P = (p_{jk})$ gives the switching probabilities of an arbitrary job and is determined by*

$$p_{jk} = \frac{\sum_{i=1}^{m} \lambda_{ij} p^{i}_{jk}}{\lambda_j} \quad \textit{for } j,k = 1,\cdots,n\,.$$

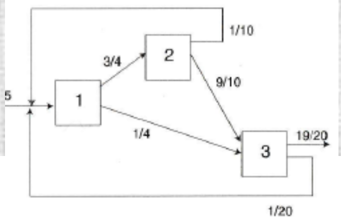
As long as there is no priority being given to specific job types, all jobs experience the same queue; therefore, the mean cycle time within Workstation *k* by Job Type *i* is given as

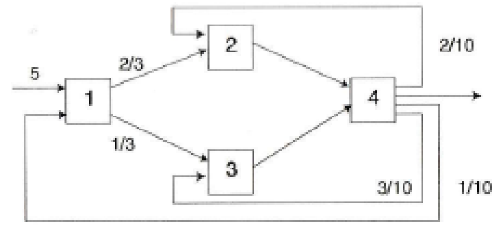
$$CT_s(i,k) = CT_q(k) + E[T_s(i,k)]\,.$$

Property 6.4. *Consider a factory of n workstations with m different job types. Assume that the external arrival rate of jobs of Type i to Workstation k is given by $\gamma_{i,k}$, and the total arrival rate of Job Type i to Workstation k is given by $\lambda_{i,k}$. Furthermore assume that the mean time spent waiting for processing in Workstation k by an arbitrary job (namely, $CT_q(k)$) has been determined. Then the mean time spent within the factory by a Type i job is given by*

$$CT_s^i = \frac{\sum_{k=1}^{n} \lambda_{ik}(CT_q(k) + E[T_s(i,k)])}{\sum_{j=1}^{n} \gamma_{ij}}$$

for $i = 1, \cdots, m$.

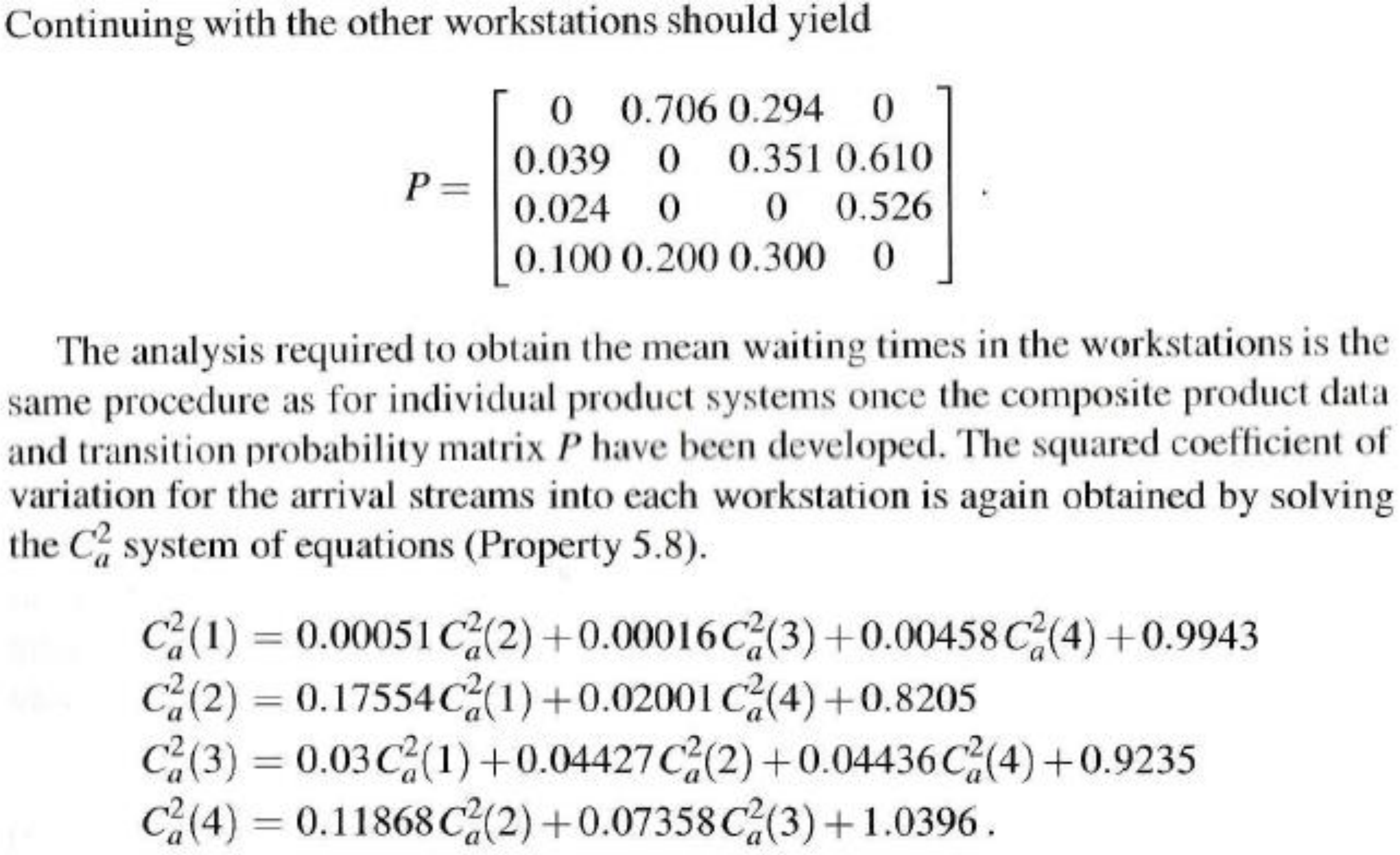




$\boldsymbol{\lambda}^1 = (5.690, 4.267, 5.263, 0)$. $\boldsymbol{\lambda}^2 = (6.25, 6.667, 5.833, 12.5)$.

Example 6.4. We now complete the analysis of the factory contained in Examples 6.1–6.3. The matrix of probabilities are obtained from Property 6.3. For example, the probability of going from Workstation 2 to Workstation 1 is determined as

$$p_{21} = \frac{\lambda_{12}p_{21}^1 + \lambda_{22}p_{21}^2}{\lambda_2} = \frac{4.267(0.1) + 6.667(0)}{10.934} = 0.039 .$$

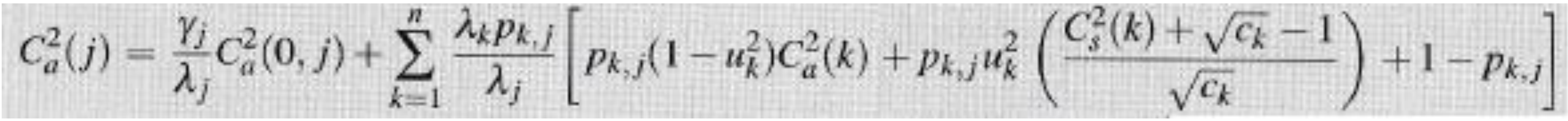


Continuing with the other workstations should yield

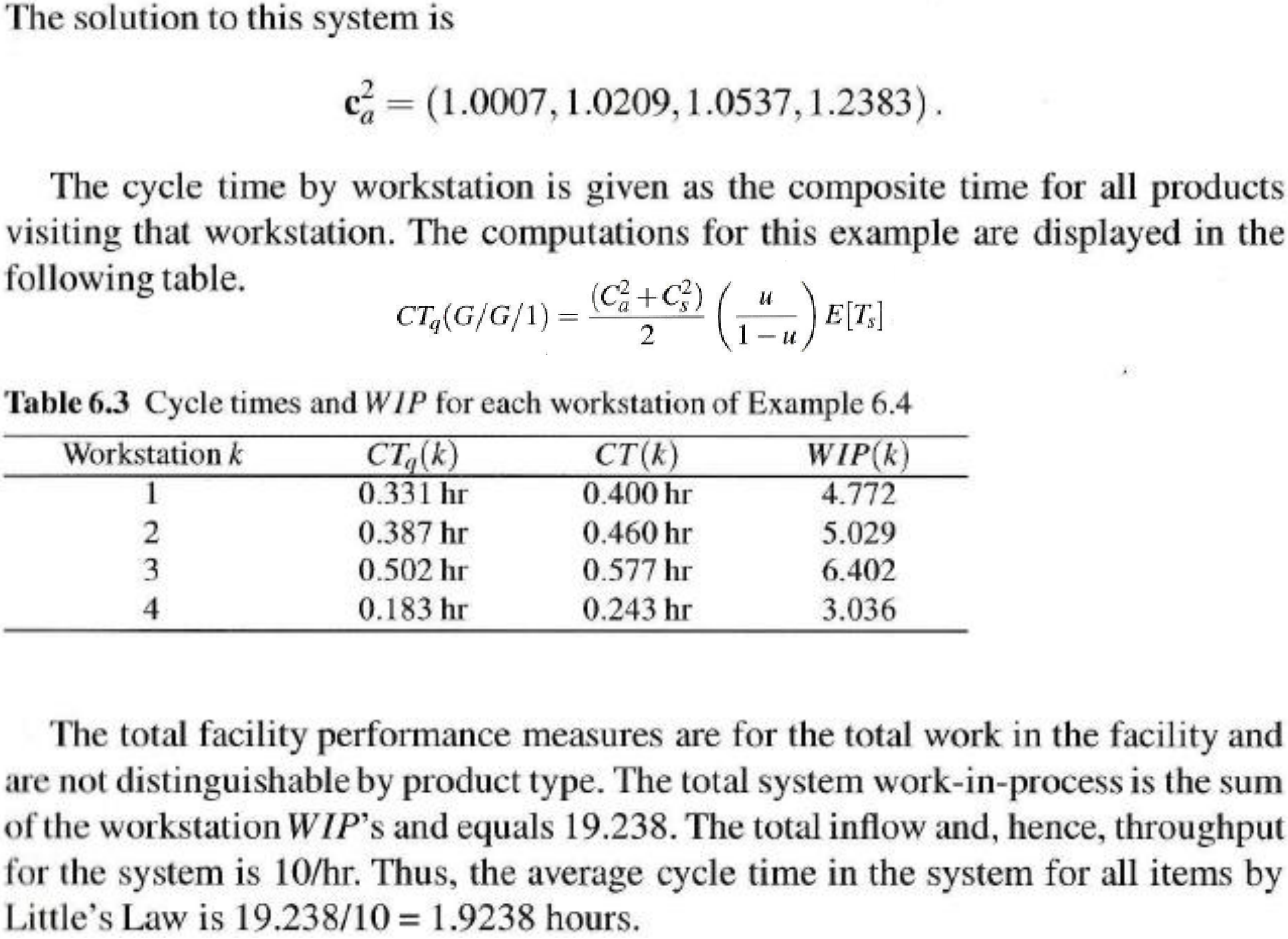
$$P = \begin{bmatrix} 0 & 0.706 & 0.294 & 0 \\ 0.039 & 0 & 0.351 & 0.610 \\ 0.024 & 0 & 0 & 0.526 \\ 0.100 & 0.200 & 0.300 & 0 \end{bmatrix} .$$

The analysis required to obtain the mean waiting times in the workstations is the same procedure as for individual product systems once the composite product data and transition probability matrix P have been developed. The squared coefficient of variation for the arrival streams into each workstation is again obtained by solving the C_a^2 system of equations (Property 5.8).

$$\begin{aligned} C_a^2(1) &= 0.00051\,C_a^2(2) + 0.00016\,C_a^2(3) + 0.00458\,C_a^2(4) + 0.9943 \\ C_a^2(2) &= 0.17554\,C_a^2(1) + 0.02001\,C_a^2(4) + 0.8205 \\ C_a^2(3) &= 0.03\,C_a^2(1) + 0.04427\,C_a^2(2) + 0.04436\,C_a^2(4) + 0.9235 \\ C_a^2(4) &= 0.11868\,C_a^2(2) + 0.07358\,C_a^2(3) + 1.0396 . \end{aligned}$$



$$C_a^2(j) = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \left[p_{k,j}(1-u_k^2)C_a^2(k) + p_{k,j}u_k^2 \left(\frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} \right) + 1 - p_{k,j} \right]$$



The solution to this system is

$$\mathbf{c}_a^2 = (1.0007, 1.0209, 1.0537, 1.2383)\,.$$

The cycle time by workstation is given as the composite time for all products visiting that workstation. The computations for this example are displayed in the following table.

$$CT_q(G/G/1) = \frac{(C_a^2 + C_s^2)}{2}\left(\frac{u}{1-u}\right)E[T_s]$$

Table 6.3 Cycle times and *WIP* for each workstation of Example 6.4

Workstation k	$CT_q(k)$	$CT(k)$	$WIP(k)$
1	0.331 hr	0.400 hr	4.772
2	0.387 hr	0.460 hr	5.029
3	0.502 hr	0.577 hr	6.402
4	0.183 hr	0.243 hr	3.036

The total facility performance measures are for the total work in the facility and are not distinguishable by product type. The total system work-in-process is the sum of the workstation *WIP*'s and equals 19.238. The total inflow and, hence, throughput for the system is 10/hr. Thus, the average cycle time in the system for all items by Little's Law is 19.238/10 = 1.9238 hours.

$$CT_s^i = \frac{\sum_{k=1}^{n} \lambda_{ik}(CT_q(k) + E[T_s(i,k)])}{\sum_{j=1}^{n} \gamma_{ij}}$$

$$\boldsymbol{\lambda}^1 = (5.690, 4.267, 5.263, 0)\,.$$

$$\boldsymbol{\lambda}^2 = (6.25, 6.667, 5.833, 12.5)\,.$$

Workstation k	$E[T_s(1,k)]$	$C_s^2(1,k)$	$E[T_s(2,k)]$	$C_s^2(2,k)$
1	1/14 hr	0.8	1/15 hr	1.33
2	1/10 hr	1.2	1/18 hr	2.00
3	1/15 hr	1.5	1/12 hr	1.50
4	—	—	0.06 hr	0.75

Table 6.3 Cycle times and *WIP* for each workstation of Example 6.4

Workstation k	$CT_q(k)$	$CT(k)$	$WIP(k)$
1	0.331 hr	0.400 hr	4.772
2	0.387 hr	0.460 hr	5.029
3	0.502 hr	0.577 hr	6.402
4	0.183 hr	0.243 hr	3.036

Property 6.4 is combined with the data of Tables 6.1 and 6.3 to produce the system mean cycle times by individual product type. For this example these computations are:

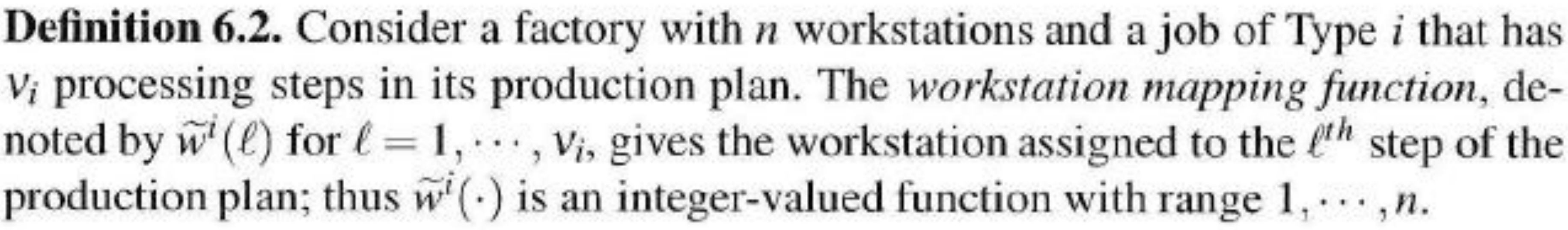
$$CT^1 = [5.690(0.3307 + 0.0714) + 4.2674(0.3870 + 0.1) + 5.2632(0.5015 + 0.0667)]\,/\,5 = 1.4714 \text{ hr}$$

$$CT^2 = [6.25(0.3307 + 0.0667) + 6.6667(0.3870 + 0.0556) + 5.8333(0.5015 + 0.0833) + 12.5(0.1828 + 0.06)]\,/\,5 = 2.3763 \text{ hr}\,.$$

These two products are produced in equal quantities, so the average cycle time for the factory is the average of these two individual product cycle times or 1.9238 hours.

1.6.5 Processing Step Modeling Paradigm

- It is necessary to keep track of not only the job location but also the visit number to the location
  - The mean and standard deviation of processing time may not be the same even if the same type of job visited the same workstation.
  - The switching probabilities may depend on both the job type and the visiting times to a workstation of the job.
- To accomplish the job location control, a data description method is used that is based on the process step that the job is undergoing.
  - List the processing steps that a job must go through during the production process.
  - The information associated with each processing step includes the workstation being visited and the processing time characteristics.



Definition 6.2. Consider a factory with n workstations and a job of Type i that has ν_i processing steps in its production plan. The *workstation mapping function*, denoted by $\widetilde{w}^i(\ell)$ for $\ell = 1, \cdots, \nu_i$, gives the workstation assigned to the ℓ^{th} step of the production plan; thus $\widetilde{w}^i(\cdot)$ is an integer-valued function with range $1, \cdots, n$.

- Consider an example shown in Table 6.5. The product flow is

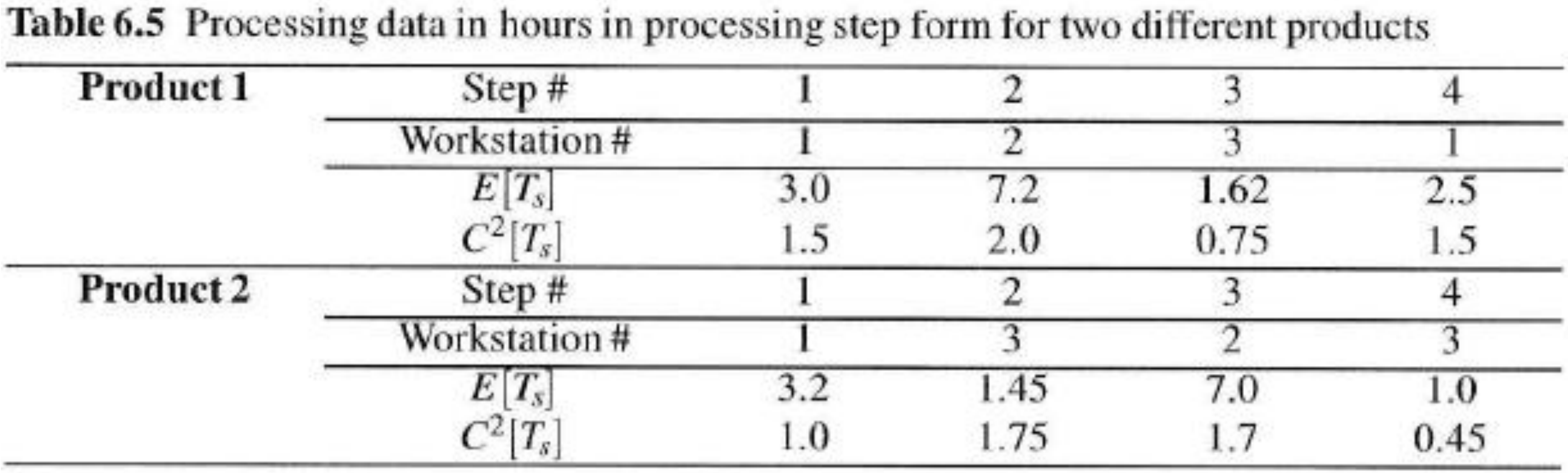
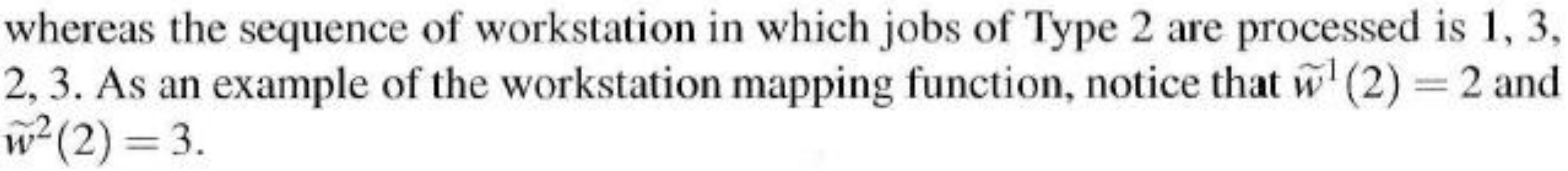


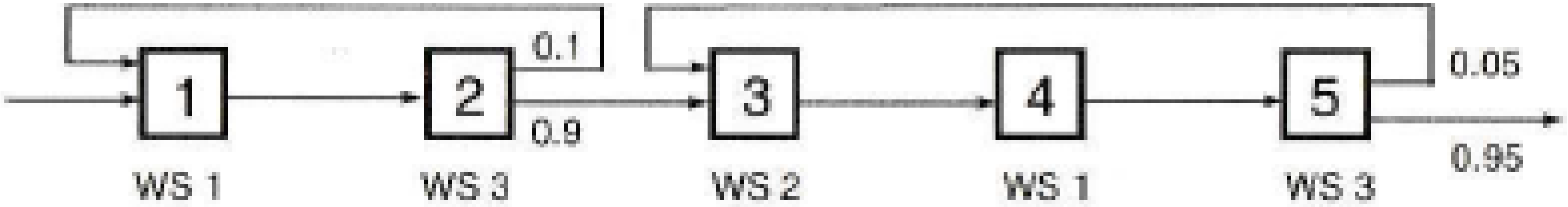
Table 6.5 Processing data in hours in processing step form for two different products

Product 1	Step #	1	2	3	4
	Workstation #	1	2	3	1
	$E[T_s]$	3.0	7.2	1.62	2.5
	$C^2[T_s]$	1.5	2.0	0.75	1.5
Product 2	Step #	1	2	3	4
	Workstation #	1	3	2	3
	$E[T_s]$	3.2	1.45	7.0	1.0
	$C^2[T_s]$	1.0	1.75	1.7	0.45



whereas the sequence of workstation in which jobs of Type 2 are processed is 1, 3, 2, 3. As an example of the workstation mapping function, notice that $\widetilde{w}^1(2) = 2$ and $\widetilde{w}^2(2) = 3$.

Definition 6.3. Consider a factory with m job types, where Job Type i has a production plan consisting of ν_i steps. The *step-wise routing matrix*, denoted by $\widetilde{P}^i$, for Job Type i is a square matrix of size $\nu_i \times \nu_i$ where $\widetilde{p}^i_{\ell,j}$ gives the probability that Job Type i will be routed to Step j after completing Step ℓ.



Example 6.5. Consider the production plan given in Table 6.6 involving a factory with three workstations. Assume that Workstations 1 and 2 are reliable but that Workstation 3 is not. There is 10% chance that jobs being processed through the third workstation for the first time (i.e., Step 2) must be returned to Workstation 1 (Step 1), and a 5% chance that jobs being processed through the third workstation for the second (i.e., Step 5) time must be returned to Workstation 2 (Step 3).

Table 6.6 Processing step paradigm for multiple visits to workstations with the data in hours

Step #	1	2	3	4	5
Workstation #	1	3	2	1	3
$E[T_s]$	3.0	2.5	3.7	4.0	3.6
$C^2[T_s]$	1.0	0.75	1.25	1.75	1.32

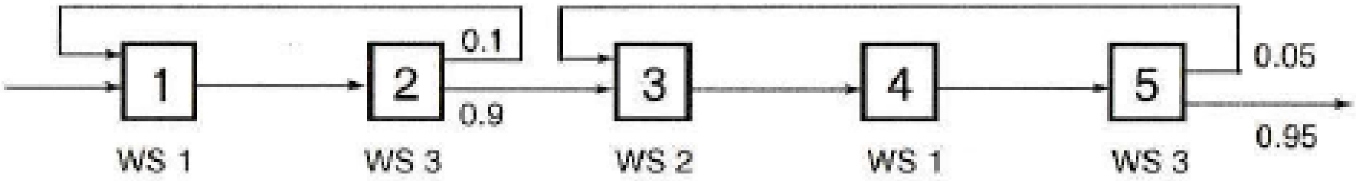


Table 6.6 Processing step paradigm for multiple visits to workstations with the data in hours

Step #	1	2	3	4	5
Workstation #	1	3	2	1	3
$E[T_s]$	3.0	2.5	3.7	4.0	3.6
$C^2[T_s]$	1.0	0.75	1.25	1.75	1.32

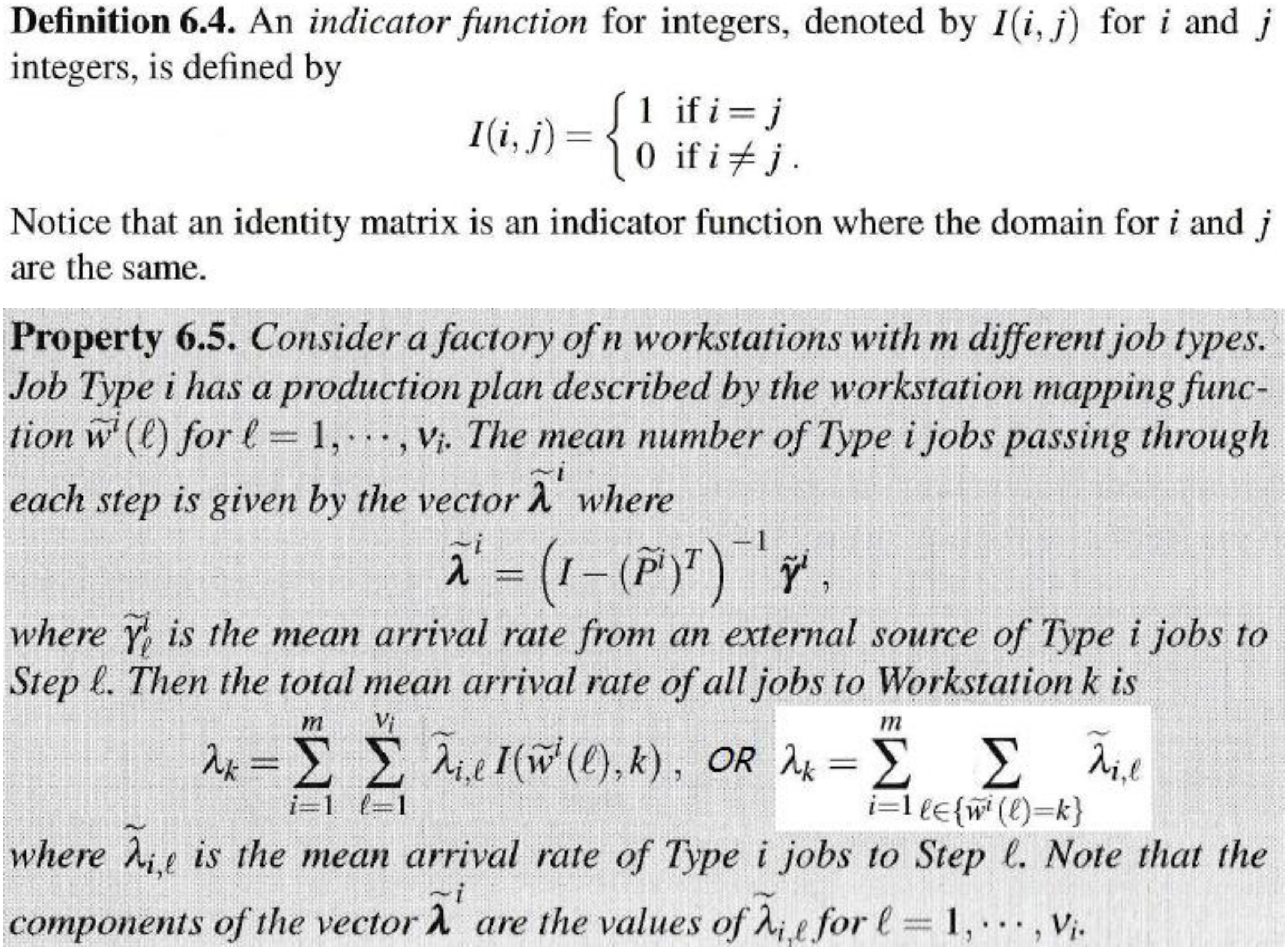
In this case, the workstation mapping function is

$$\widetilde{w}^1(1) = 1,\ \widetilde{w}^1(2) = 3,\ \widetilde{w}^1(3) = 2,\ \widetilde{w}^1(4) = 1,\ \widetilde{w}^1(5) = 3\ ,$$

and the step-wise routing matrix is given by

$$\widetilde{P}^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.05 & 0 & 0 \end{bmatrix} .$$

1.6.5.1 Service Time Characteristics



Definition 6.4. An *indicator function* for integers, denoted by $I(i,j)$ for i and j integers, is defined by

$$I(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j\,. \end{cases}$$

Notice that an identity matrix is an indicator function where the domain for i and j are the same.

Property 6.5. *Consider a factory of n workstations with m different job types. Job Type i has a production plan described by the workstation mapping function $\widetilde{w}^i(\ell)$ for $\ell = 1, \cdots, \nu_i$. The mean number of Type i jobs passing through each step is given by the vector $\widetilde{\boldsymbol{\lambda}}^i$ where*

$$\widetilde{\boldsymbol{\lambda}}^i = \left(I - (\widetilde{P}^i)^T\right)^{-1} \widetilde{\boldsymbol{\gamma}}^i\,,$$

where $\widetilde{\gamma}^i_\ell$ is the mean arrival rate from an external source of Type i jobs to Step ℓ. Then the total mean arrival rate of all jobs to Workstation k is

$$\lambda_k = \sum_{i=1}^{m} \sum_{\ell=1}^{\nu_i} \widetilde{\lambda}_{i,\ell}\, I(\widetilde{w}^i(\ell), k)\,, \quad \text{OR} \quad \lambda_k = \sum_{i=1}^{m} \sum_{\ell \in \{\widetilde{w}^i(\ell) = k\}} \widetilde{\lambda}_{i,\ell}$$

where $\widetilde{\lambda}_{i,\ell}$ is the mean arrival rate of Type i jobs to Step ℓ. Note that the components of the vector $\widetilde{\boldsymbol{\lambda}}^i$ are the values of $\widetilde{\lambda}_{i,\ell}$ for $\ell = 1, \cdots, \nu_i$.

We let the random variable $\widetilde{T}_s(i,\ell)$ denote the processing time for Job Type i during the ℓ^{th} step of its production plan. The mean service time for Job Type i during Step ℓ is denoted by $E[\widetilde{T}_s(i,\ell)]$ and this occurs at the workstation designated by $\widetilde{w}^i(\ell)$. Likewise, the squared coefficient of variation of the service time is given by $\widetilde{C}_s^2(i,\ell)$. With these definitions, the workload and utilization for Workstation k

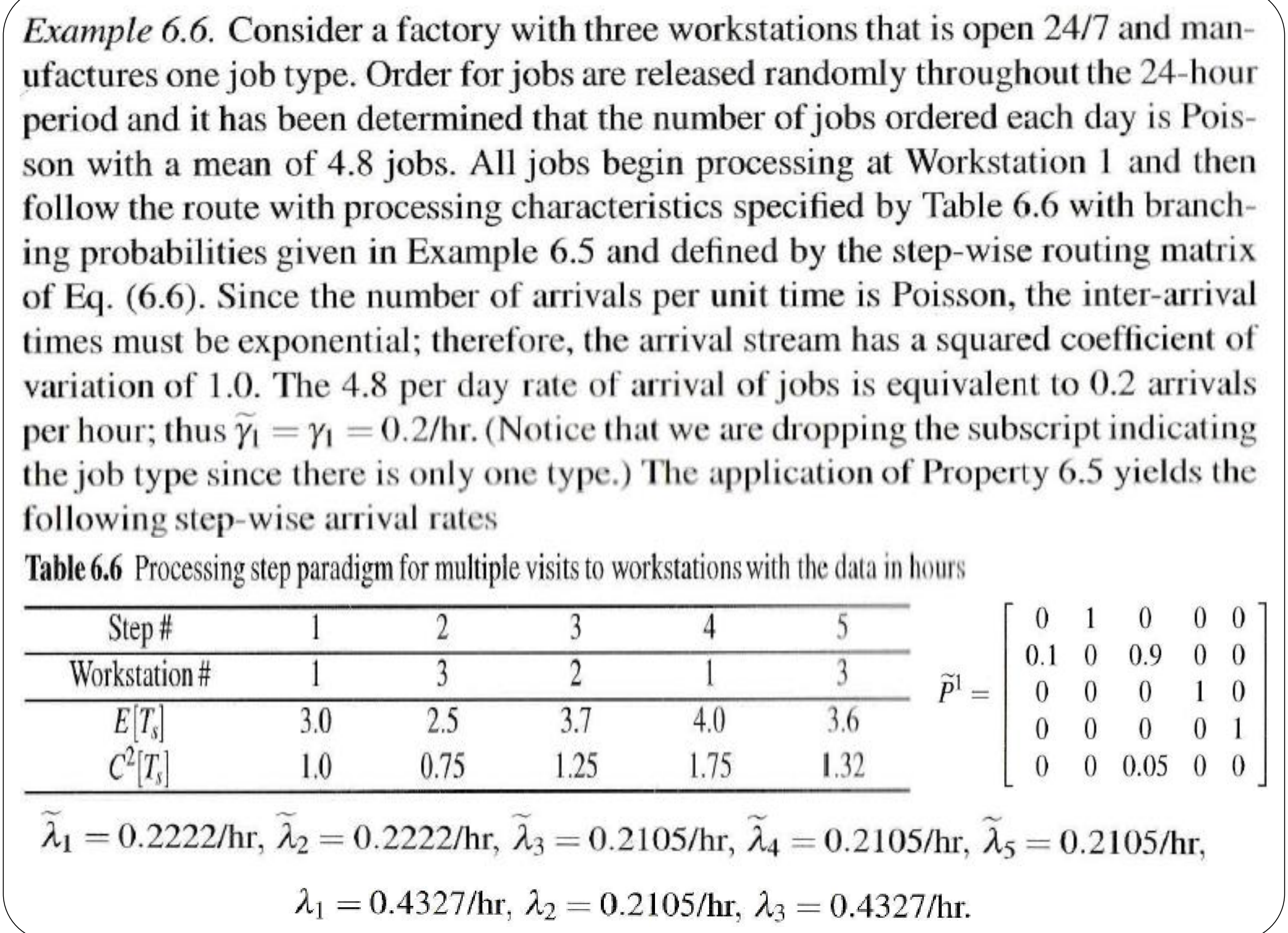
$$u_k = \frac{WL_k}{c_k} = \left(\sum_{i=1}^{m} \sum_{\ell=1}^{v_i} \widetilde{\lambda}_{i,\ell} E[\widetilde{T}_s(i,\ell)]\, I(\widetilde{w}^i(\ell),k) \right) / \, c_k \ ,$$

where c_k is the number of identical processors available at Workstation k to handle the workload, m is the number of job types, and v_i is the number of production steps for Job Type i.

The service time characteristics for Workstation k are also given similarly

$$E[T_s(k)] = \sum_{i=1}^{m} \sum_{\ell=1}^{v_i} \frac{\widetilde{\lambda}_{i,\ell}}{\lambda_k} E[\widetilde{T}_s(i,\ell)]\, I(\widetilde{w}^i(\ell),k) = \frac{WL_k}{\lambda_k} \ ,$$

$$C_s^2(k) = \frac{\sum_{i=1}^{m} \sum_{\ell=1}^{v_i} (\widetilde{\lambda}_{i,\ell}/\lambda_k)\, E[\widetilde{T}_s(i,\ell)]^2 (1 + \widetilde{C}_s^2(i,\ell))\, I(\widetilde{w}^i(\ell),k)}{E[T_s(k)]^2} - 1 \ .$$



Example 6.6. Consider a factory with three workstations that is open 24/7 and manufactures one job type. Order for jobs are released randomly throughout the 24-hour period and it has been determined that the number of jobs ordered each day is Poisson with a mean of 4.8 jobs. All jobs begin processing at Workstation 1 and then follow the route with processing characteristics specified by Table 6.6 with branching probabilities given in Example 6.5 and defined by the step-wise routing matrix of Eq. (6.6). Since the number of arrivals per unit time is Poisson, the inter-arrival times must be exponential; therefore, the arrival stream has a squared coefficient of variation of 1.0. The 4.8 per day rate of arrival of jobs is equivalent to 0.2 arrivals per hour; thus $\widetilde{\gamma}_1 = \gamma_1 = 0.2$/hr. (Notice that we are dropping the subscript indicating the job type since there is only one type.) The application of Property 6.5 yields the following step-wise arrival rates

Table 6.6 Processing step paradigm for multiple visits to workstations with the data in hours

Step #	1	2	3	4	5
Workstation #	1	3	2	1	3
$E[T_s]$	3.0	2.5	3.7	4.0	3.6
$C^2[T_s]$	1.0	0.75	1.25	1.75	1.32

$$\widetilde{P}^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.05 & 0 & 0 \end{bmatrix}$$

$$\widetilde{\lambda}_1 = 0.2222/\text{hr},\ \widetilde{\lambda}_2 = 0.2222/\text{hr},\ \widetilde{\lambda}_3 = 0.2105/\text{hr},\ \widetilde{\lambda}_4 = 0.2105/\text{hr},\ \widetilde{\lambda}_5 = 0.2105/\text{hr},$$

$$\lambda_1 = 0.4327/\text{hr},\ \lambda_2 = 0.2105/\text{hr},\ \lambda_3 = 0.4327/\text{hr}.$$

The workload calculations for the three workstations are

$$WL_1 = 0.2222 \times 3.0 + 0.2105 \times 4.0 = 1.5086$$

$$WL_2 = 0.2105 \times 3.7 = 0.7789$$

$$WL_3 = 0.2222 \times 2.5 + 0.2105 \times 3.6 = 1.3140\,.$$

For a steady-state to exist, the number of machines at each workstation must be strictly greater than the workload; therefore, there must be at least two machines for Workstations 1 and 3 and one machine at Workstation 2. Assuming the minimum requirements, the workstation utilization vector is (75.4%, 78.0%, 65.7%).

The service time characteristics for Workstation 1 are calculated as

$$E[T_s(1)] = \frac{1.5086}{0.4327} = 3.486 \text{ and}$$

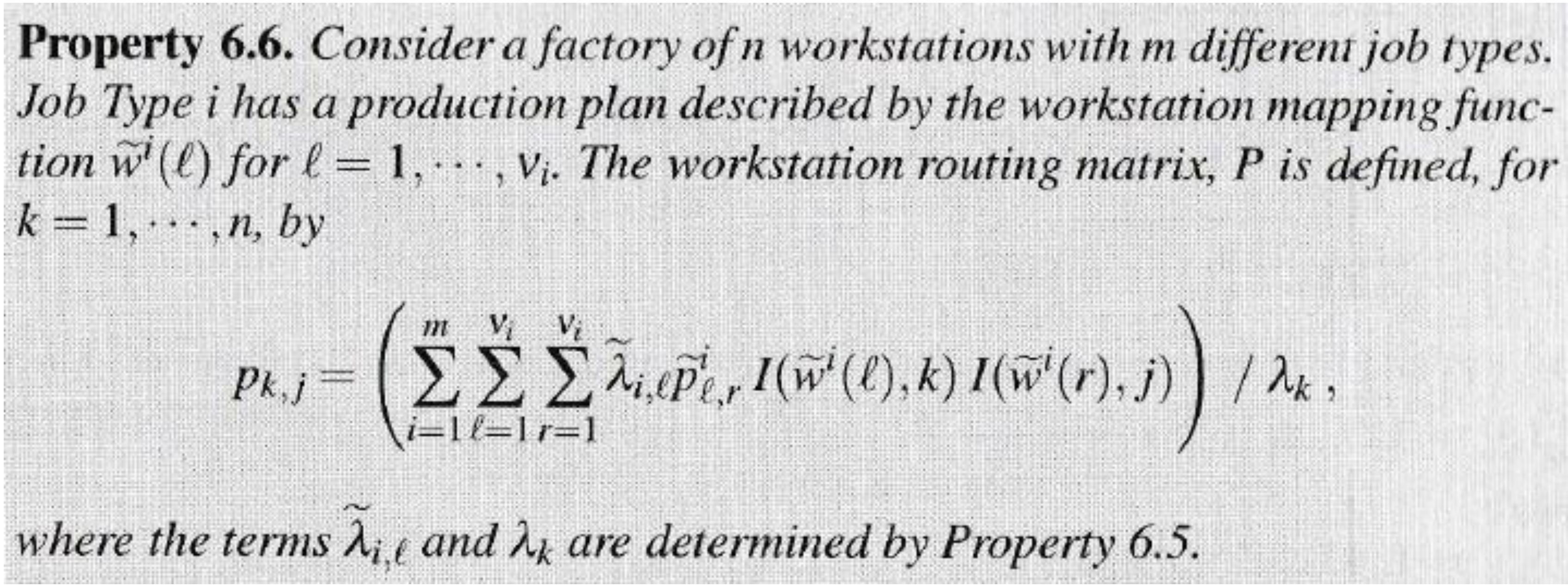
$$C^2[T_s(1)] = \frac{(0.2222/0.4327)(3^2)(1+1) + (0.2105/0.4327)(4^2)(1+1.75)}{3.486^2} - 1 = 1.522$$

Table 6.7 The composite processing data for Example 6.6

Workstation k	c_k	u_k	$E[T_s(k)]$	$C_s^2(k)$
1	2	0.754	3.486 hr	1.522
2	1	0.780	3.700 hr	1.252
3	2	0.657	3.037 hr	1.195

1.6.5.2 Performance Measures

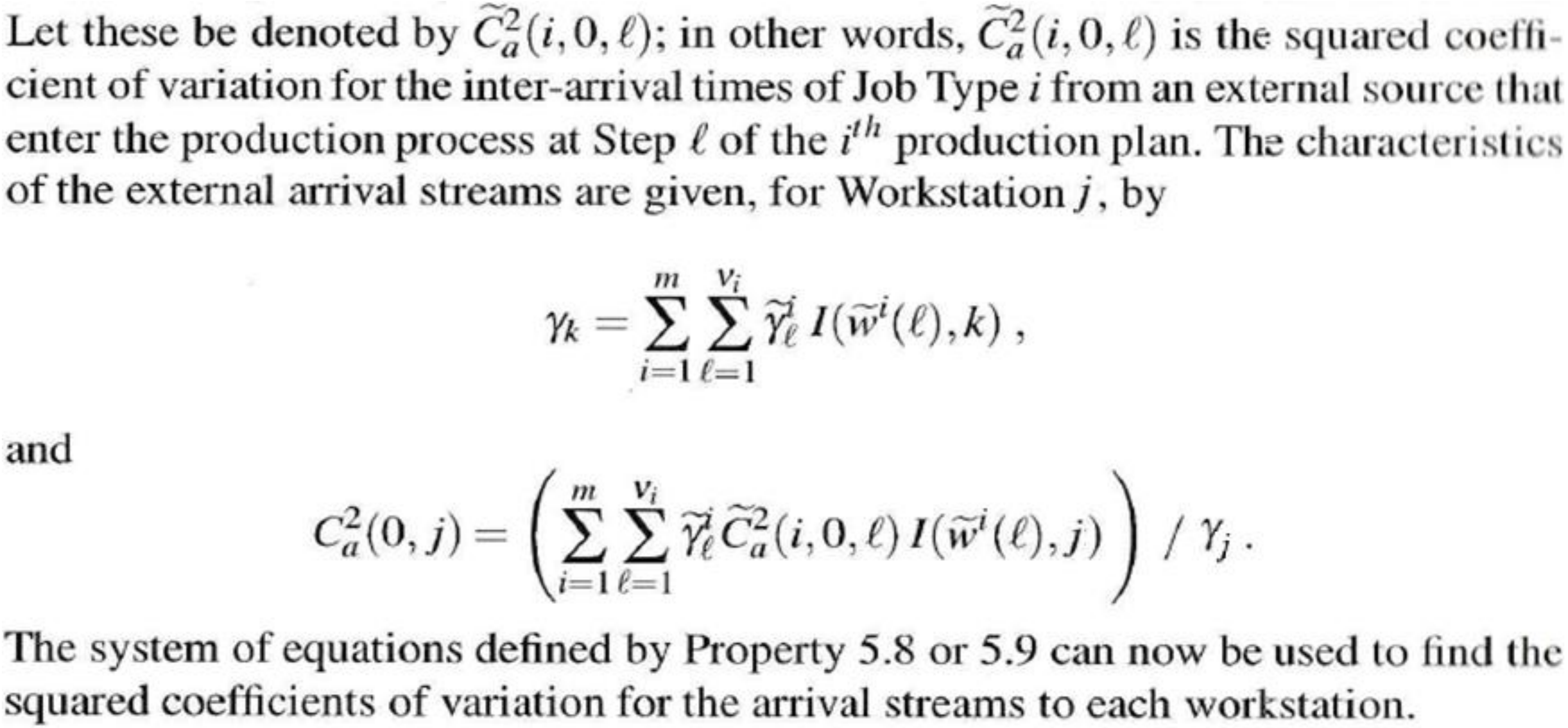
- Complete the factory analysis
  - Mean and squared coefficients of variation for the processing times of a workstation
  - Mean and squared coefficients of variation for the arrival streams to each work station
- To obtain the system of equations that define these terms, the factory with multiple routing schemes will be converted to a similar factory with probabilistic routing by the following route matrix.



Property 6.6. *Consider a factory of n workstations with m different job types. Job Type i has a production plan described by the workstation mapping function $\widetilde{w}^i(\ell)$ for $\ell = 1, \cdots, \nu_i$. The workstation routing matrix, P is defined, for $k = 1, \cdots, n$, by*

$$p_{k,j} = \left(\sum_{i=1}^{m} \sum_{\ell=1}^{\nu_i} \sum_{r=1}^{\nu_i} \widetilde{\lambda}_{i,\ell} \widetilde{p}^i_{\ell,r} \, I(\widetilde{w}^i(\ell),k) \, I(\widetilde{w}^i(r),j) \right) / \lambda_k ,$$

where the terms $\widetilde{\lambda}_{i,\ell}$ and λ_k are determined by Property 6.5.



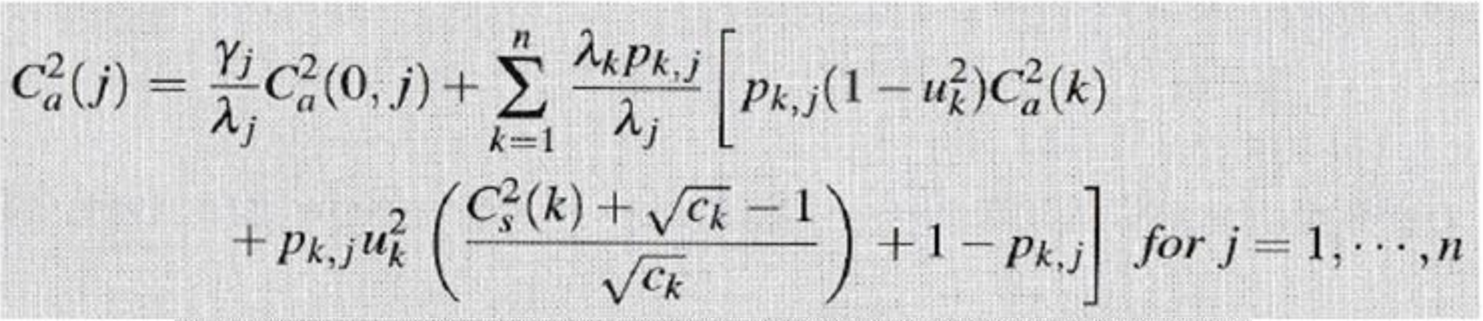
Let these be denoted by $\widetilde{C}_a^2(i,0,\ell)$; in other words, $\widetilde{C}_a^2(i,0,\ell)$ is the squared coefficient of variation for the inter-arrival times of Job Type i from an external source that enter the production process at Step ℓ of the i^{th} production plan. The characteristics of the external arrival streams are given, for Workstation j, by

$$\gamma_k = \sum_{i=1}^{m} \sum_{\ell=1}^{v_i} \widetilde{\gamma}_\ell^i \, I(\widetilde{w}^i(\ell), k) \, ,$$

and

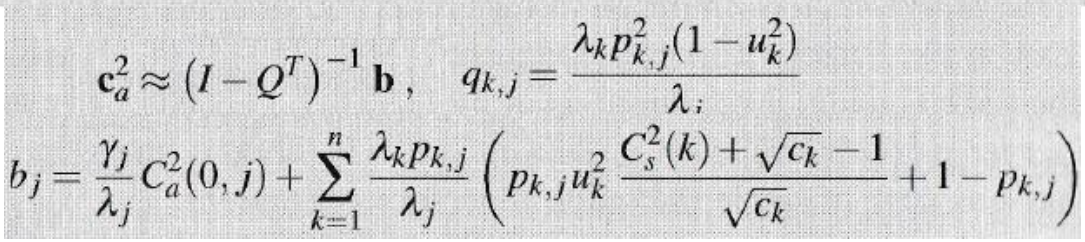
$$C_a^2(0,j) = \left(\sum_{i=1}^{m} \sum_{\ell=1}^{v_i} \widetilde{\gamma}_\ell^i \, \widetilde{C}_a^2(i,0,\ell) \, I(\widetilde{w}^i(\ell), j) \right) \Big/ \, \gamma_j \, .$$

The system of equations defined by Property 5.8 or 5.9 can now be used to find the squared coefficients of variation for the arrival streams to each workstation.



$$C_a^2(j) = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \left[p_{k,j}(1-u_k^2) C_a^2(k) \right.$$
$$\left. + \, p_{k,j} u_k^2 \left(\frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} \right) + 1 - p_{k,j} \right] \; for \; j = 1, \cdots, n$$

or



$$\mathbf{c}_a^2 \approx \left(I - Q^T \right)^{-1} \mathbf{b} \, , \quad q_{k,j} = \frac{\lambda_k p_{k,j}^2 (1-u_k^2)}{\lambda_j}$$

$$b_j = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \left(p_{k,j} u_k^2 \, \frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} + 1 - p_{k,j} \right)$$

Example 6.7. Example 6.6 can now be completed (Fig. 6.1). The associated average product routing matrix for the three workstations obtained from Property 6.6 is

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0.538 & 0 \end{bmatrix} .$$

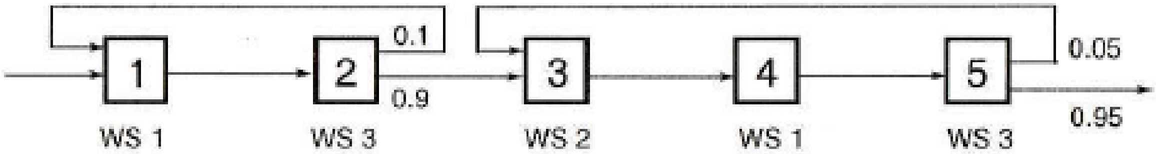


Table 6.6 Processing step paradigm for multiple visits to workstations with the data in hours

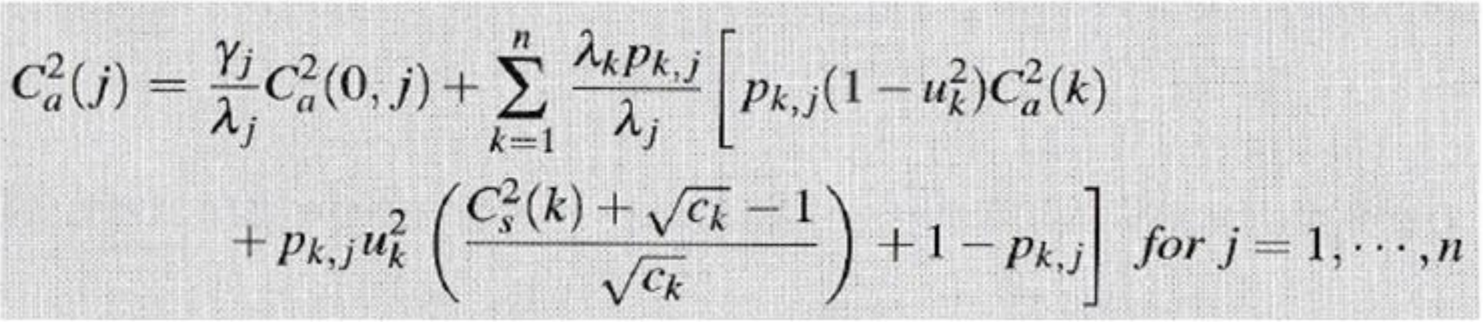
Step #	1	2	3	4	5
Workstation #	1	3	2	1	3
$E[T_s]$	3.0	2.5	3.7	4.0	3.6
$C^2[T_s]$	1.0	0.75	1.25	1.75	1.32

$$\widetilde{P}^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.05 & 0 & 0 \end{bmatrix}$$

$$\widetilde{\lambda}_1 = 0.2222/\text{hr},\ \widetilde{\lambda}_2 = 0.2222/\text{hr},\ \widetilde{\lambda}_3 = 0.2105/\text{hr},\ \widetilde{\lambda}_4 = 0.2105/\text{hr},\ \widetilde{\lambda}_5 = 0.2105/\text{hr},$$

$$\lambda_1 = 0.4327/\text{hr},\ \lambda_2 = 0.2105/\text{hr},\ \lambda_3 = 0.4327/\text{hr}.$$

$$p_{k,j} = \left(\sum_{i=1}^{m} \sum_{\ell=1}^{v_i} \sum_{r=1}^{v_i} \widetilde{\lambda}_{i,\ell} \widetilde{p}^{i}_{\ell,r} I(\widetilde{w}^{i}(\ell), k)\, I(\widetilde{w}^{i}(r), j) \right) / \lambda_k$$



$$C_a^2(j) = \frac{\gamma_j}{\lambda_j} C_a^2(0,j) + \sum_{k=1}^{n} \frac{\lambda_k p_{k,j}}{\lambda_j} \Bigg[p_{k,j}(1-u_k^2) C_a^2(k) + p_{k,j} u_k^2 \left(\frac{C_s^2(k) + \sqrt{c_k} - 1}{\sqrt{c_k}} \right) + 1 - p_{k,j} \Bigg] \quad \textit{for } j = 1, \cdots, n$$

The system of equations for computing the coefficients of variation for the average product arrival streams at each workstation is

$$C_a^2(1) = \frac{0.2}{0.4327}(1) + \frac{0.2105}{0.4327}\left[\left(1-u_2^2\right) C_a^2(2) + u_2^2 C_s^2(2)\right] = 0.1905 C_a^2(2) + 0.8328$$

$$C_a^2(2) = \frac{0.4327 \times 0.538}{0.2105} \times \Bigg[0.538\left(1-u_3^2\right) C_a^2(3) + 0.538 u_3^2 \left(\frac{C_s^2(3) + \sqrt{2} - 1}{\sqrt{2}} \right) + 1 - 0.538 \Bigg] = 0.3382 C_a^2(3) + 0.8032$$

$$C_a^2(3) = \left(1-u_1^2\right) C_a^2(1) + u_1^2 \left(\frac{C_s^2(1) + \sqrt{2} + -1}{\sqrt{2}} \right) = 0.4315 C_s^2(1) + 0.7784\,.$$

The solution to this linear system of equations $\mathbf{c}_a^2 = (1.066, 1.222, 1.238)\,.$

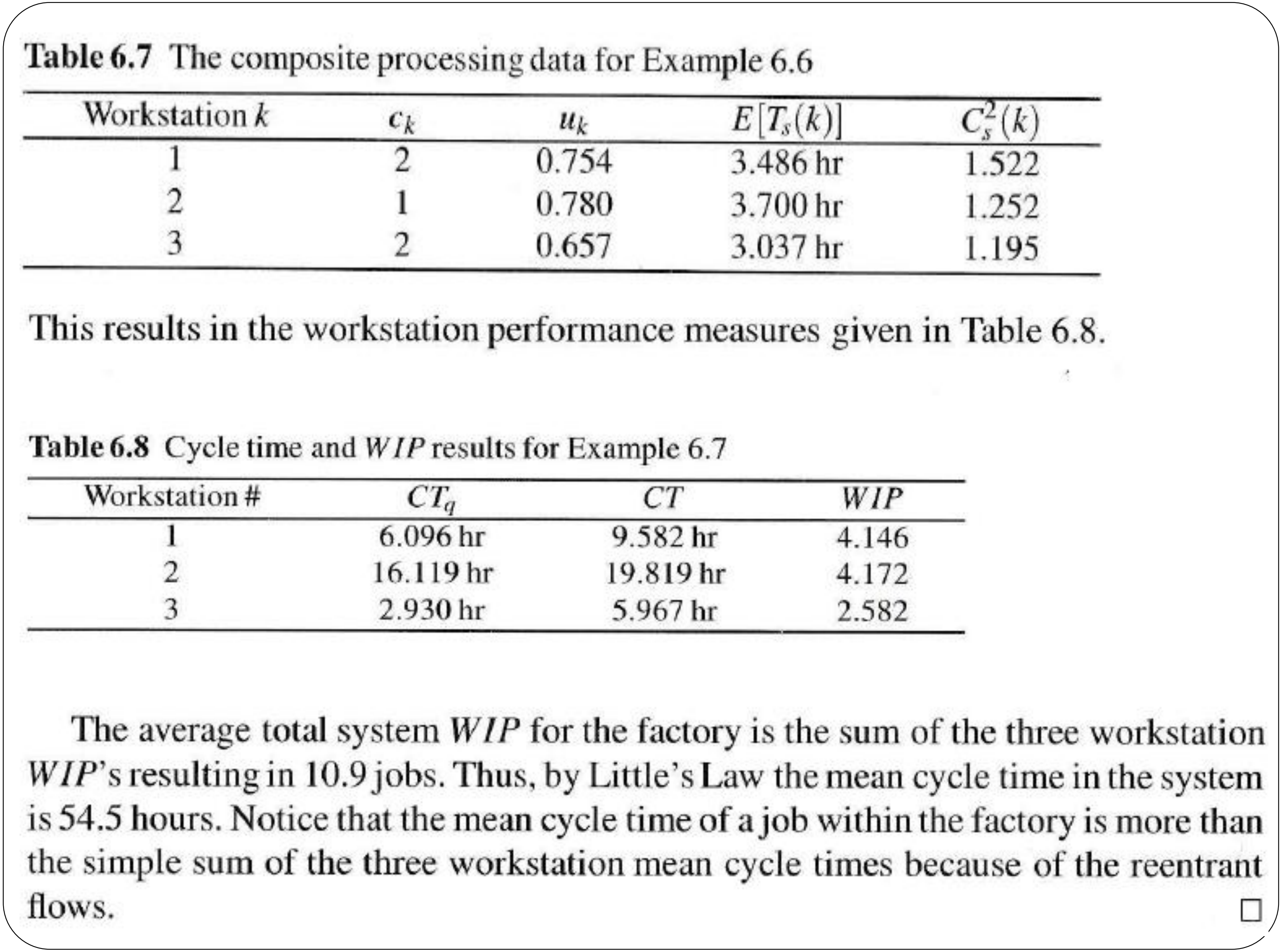


Table 6.7 The composite processing data for Example 6.6

Workstation k	c_k	u_k	$E[T_s(k)]$	$C_s^2(k)$
1	2	0.754	3.486 hr	1.522
2	1	0.780	3.700 hr	1.252
3	2	0.657	3.037 hr	1.195

This results in the workstation performance measures given in Table 6.8.

Table 6.8 Cycle time and *WIP* results for Example 6.7

Workstation #	CT_q	CT	WIP
1	6.096 hr	9.582 hr	4.146
2	16.119 hr	19.819 hr	4.172
3	2.930 hr	5.967 hr	2.582

The average total system *WIP* for the factory is the sum of the three workstation *WIP*'s resulting in 10.9 jobs. Thus, by Little's Law the mean cycle time in the system is 54.5 hours. Notice that the mean cycle time of a job within the factory is more than the simple sum of the three workstation mean cycle times because of the reentrant flows. □