UNIVERSITY OF DUBLIN
TRINITY COLLEGE

FACULTY OF ENGINEERING & SYSTEMS SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

BA (Mod.) Computer Science **Trinity Term 2000**
Degree Examination

4BA2- Systems Modelling

Wednesday 24th May **Goldsmith Hall** **14.00 - 17.00**

Professor Francis Neelamkavil, Mr. Donal O'Mahony, Dr. Tony Redmond

Answer FIVE questions, at least one from each section
Please use separate answer books for each section.
(Queuing Tables attached to paper)

SECTION A

1. Using appropriate diagrams or simple examples, explain the CA (Cellular Automata) and GA (Genetic Algorithm) approaches to problem solving. Comment on the merits and de-merits of these two methods.

2. Define the terms "modelling" and "simulation".
With the help of a block diagram, explain a general methodology for iterative modelling and simulation of systems.
What are the limitations of simulation?

SECTION B

3. Describe the underlying structure of the links used to transport telephony traffic between exchanges in the public switched telephone network. Explain how owners of such network infrastructure or companies who lease capacity can use these facilities to provide an array of services to users.

4. Give an outline of the architecture and the major protocols involved in realizing the Internet E-mail service. What elements would need to be added to allow Internet E-mail to be used as a store-and-forward service for business-to-business E-commerce. Are there likely to be any non-technical obstacles to the uptake of this technology?

SECTION C

5. **a.** What is meant by Streeter's scaling effect and why is it important?

   b. A small business is considering either a configuration of 2 machines (with separate queues) of speed $\mu/2$ or a configuration of 1 machine of speed μ. Use the attached queueing theory formulae to estimate W2/W1 where W2 is the system response time in the 2 machine case and W1 is the response time in the 1 machine case. Which configuration is better and why?

   c. The company is also considering a duplex (i.e. 2 machines with 1 queue) configuration of 2 machines of speed $\mu/2$ versus the 1 machine of speed μ configuration. Derive a formula for W2/W1 and hence calculate values of it for $\rho = 0.2, 0.4, 0.6$ and 0.8. Which cofiguration is better and why?

6. **a.** Moore's Law suggests that processor power is improving at the rate of doubling every eighteen months. It has been postulated for some time that there is a "brick wall" to this improvement due to physical reasons such as possible mask size, physical problems due to smallness including capacitance problems, heat dissipation etc. Discuss this topic briefly in particular focussing on:

   i. the evidence to support this hypothesis and
   ii. if so, whether it may become a reality within the next 10 years.

   b. It has been postulated for quite some years that the CRT screen and the Mainframe will disappear. While the CRT screen is about to disappear for computer displays, the mainframe still persists. Discuss briefly as to why or why not the mainframe will continue for at least 10 years in large enterprises.

7. **a.** Give 3 queueing rules of thumb and briefly explain their importance in each case.
   b. Give briefly 6 underlying assumptions of Buzen's Central Server model
   c. Give 6 extensions of the model explaining briefly their importance in each case.
   d. Give 2 definitions of a balanced situation in a queueing network.

Appendix C

QUEUEING THEORY DEFINITIONS AND FORMULAS

In Figs. 5.1.1 and 5.1.2, reproduced from Chapter 5, we indicate the elements and random variables used in queueing theory models. Table 1 is a compendium of the queueing theory definitions and notation used in this book. The remainder of Appendix C consists of tables of queueing theory formulas for the most useful models and figures to help with the calculations. APL functions are displayed in Appendix B to implement the formulas for most of the queueing models.

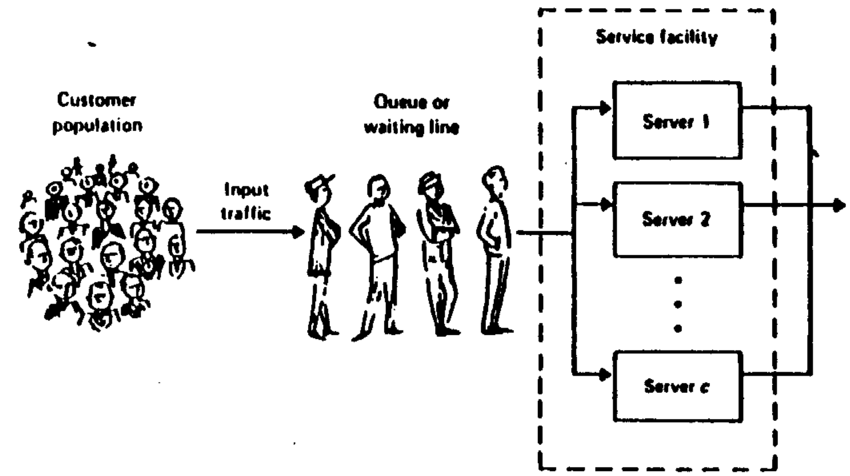

Fig. 5.1.1 Elements of a queueing system.

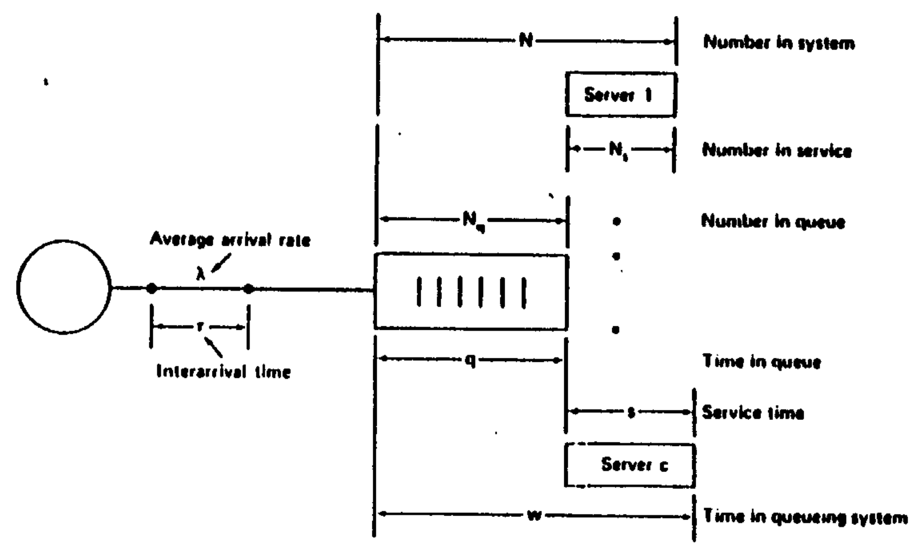

Fig. 5.1.2 Some random variables used in queueing theory models.

TABLE 1
Queueing Theory Notation and Definitions

$A(t)$	Distribution of interarrival time. $A(t) = P[\tau \leq t]$.
$B(c, u)$	Erlang's B formula or the probability all c servers are busy in an $M/M/c/c$ queueing system.
$C(c, u)$	Erlang's C formula or the probability all c servers are busy in an $M/M/c$ queueing system.
c	Symbol for the number of servers in the service facility of a queueing system.
D	Symbol for constant (deterministic) interarrival or service time distribution.
$E[N]$	Expected (average or mean) number of customers in the steady state queueing system. The letter L is also used for $E[N]$.
$E[N_q]$	Expected (average or mean) number of customers in the queue (waiting line) when the system is in the steady state. The symbol L_q is also used for $E[N_q]$.
$E[N_s]$	Expected (average or mean) number of customers receiving service when the system is in the steady state.
$E[q]$	Expected (average or mean) queueing time (does not include service time) when the system is in the steady state. The symbol W_q is also used for $E[q]$.
$E[s]$	Expected (average or mean) service time for one customer. The symbol W_s is also used for $E[s]$.
$E[\tau]$	Expected (average or mean) interarrival time. $E[\tau] = 1/\lambda$, where λ is average arrival rate.
$E[w]$	Expected (average or mean) waiting time in the system (this includes both queueing time and service time) when the system is in the steady state. The letter W is also used for $E[w]$.
E_k	Symbol for Erlang-k distribution of interarrival or service time.
$E[N_q \mid N_q > 0]$	Expected (average or mean) queue length of nonempty queues when the system is in the steady state.
$E[q \mid q > 0]$	Expected (average or mean) waiting time in queue for customers delayed when the system is in the steady state. Same as $W_{q \mid q > 0}$.
FCFS	Symbol for "first come, first served." queue discipline.
FIFO	Symbol for "first in, first out." queue discipline which is identical with FCFS.
G	Symbol for general probability distribution of service time. Independence usually assumed.
GI	Symbol for general independent interarrival time distribution.
K	Maximum number allowed in queueing system, including both those waiting for service and those receiving service. Also size of population in finite population models.
L	$E[N]$, expected (average or mean) number in the queueing system when the system is in the steady state.
$\ln(\cdot)$	The natural logarithm function or the logarithm to the base e.
L_q	$E[N_q]$, expected (average or mean) number in the queue, not including those in service, for steady state system.
LCFS	Symbol for "last come, first served." queue discipline.
LIFO	Symbol for "last in, first out." queue discipline which is identical to LCFS.
λ	Average (mean) arrival rate to queueing system. $\lambda = 1/E[\tau]$, where $E[\tau]$ = average interarrival time.

λ_T	Average throughput of a computer system measured in jobs or interactions per unit time.
M	Symbol for exponential interarrival or service time distribution.
μ	Average (mean) service rate per server. Average service rate $\mu = 1/E[s]$, where $E[s]$ is the average (mean) service time.
N	Random variable describing number in queueing system when system is in the steady state.
N_q	Random variable describing number of customers in the steady state queue.
N_s	Random variable describing number of customers receiving service when the system is in the steady state.
O	Operating time of a machine in the machine repair queueing model (Sections 5.2.6 and 5.2.7). O is the time a machine remains in operation after repair before repair again is necessary.
$p_n(t)$	Probability that there are n customers in the queueing system at time t.
p_n	Steady state probability that there are n customers in the queueing system.
PRI	Symbol for priority queueing discipline.
PS	Abbreviation for "processor-sharing queue discipline." See Section 6.2.1.
$\pi_q(r)$	Symbol for rth percentile queueing time; that is, the queueing time that r percent of the customers do not exceed.
$\pi_w(r)$	Symbol for rth percentile waiting time in the system; that is, the time in the system (queueing time plus service time) that r percent of the customers do not exceed.
q	Random variable describing the time a customer spends in the queue (waiting line) before receiving service.
RSS	Symbol for queue discipline with "random selection for service."
ρ	Server utilization = traffic intensity/c = $\lambda E[s]/c = (\lambda/\mu)/c$. The probability that any particular server is busy.
s	Random variable describing service time for one customer.
SIRO	Symbol for queue discipline, "service in random order" which is identical with RSS. It means that each waiting customer has the same probability of being served next.
τ	Random variable describing interarrival time.
u	Traffic intensity = $E[s]/E[\tau] = \lambda E[s] = \lambda/\mu$. Unit of measure is the erlang.
w	Random variable describing the total time a customer spends in the queueing system, including both service time and time spent queueing for service.
$W(t)$	Distribution function for w. $W(t) = P[w \le t]$.
W	$E[w]$, expected (average or mean) time in the steady state system.
$W_q(t)$	Distribution function for time in the queue. $W_q(t) = P[q \le t]$.
W_q	$E[q]$, expected (average or mean) time in the queue (waiting line), excluding service time, for steady state system.
$W_{q\|q>0}$	Expected (average or mean) queueing time for those who must queue. Same as $E[q \mid q > 0]$.
$W_s(t)$	Distribution function for service time. $W_s(t) = P[s \le t]$.
W_s	$E[s]$, expected (average or mean) service time, $1/\mu$.

TABLE 2

Relationships Between Random Variables of Queueing Theory Models

$u = E[s]/E[\tau] = \lambda E[s] = \lambda/\mu$	Traffic intensity in erlangs.
$\rho = u/c = \lambda E[s]/c = \lambda/c\mu$	Server utilization. The probability any particular server is busy.
$w = q + s$	Total waiting time in the system, including waiting in queue and service time.
$W = E[w] = E[q] + E[s] = W_q + W_s$	Average total waiting time in the steady state system.
$N = N_q + N_s$	Number of customers in the steady state system.
$L = E[N] = E[N_q] + E[N_s] = \lambda E[w] = \lambda W$	Average number of customers in the steady state system. $L = \lambda W$ is known as "Little's formula."
$L_q = E[N_q] = \lambda E[q] = \lambda W_q$	Average number in the queue for service for steady state system. $L_q = \lambda W_q$ is also called "Little's formula."

TABLE 3

Steady State Formulas for M/M/1 Queueing System

$$p_n = P[N = n] = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots.$$

$$P[N \geq n] = \sum_{k=n}^{\infty} p_k = \rho^n, \qquad n = 0, 1, 2, \ldots.$$

$$L = E[N] = \rho/(1 - \rho), \qquad \sigma_N^2 = \rho/(1 - \rho)^2.$$

$$L_q = E[N_q] = \rho^2/(1 - \rho), \qquad \sigma_{N_q}^2 = \rho^2(1 + \rho - \rho^2)/(1 - \rho)^2.$$

$$E[N_q | N_q > 0] = 1/(1 - \rho), \qquad \mathrm{Var}[N_q | N_q > 0] = \rho/(1 - \rho)^2.$$

$$W(t) = P[w \leq t] = 1 - e^{-t/W}, \qquad P[w > t] = e^{-t/W}.$$

$$W = E[w] = E[s]/(1 - \rho), \qquad \sigma_w = W.$$

$$\pi_w(90) = W \ln 10 \approx 2.3W, \qquad \pi_w(95) = W \ln 20 \approx 3W.$$

$$\pi_w(r) = W \ln [100/(100 - r)].$$

$$W_q(t) = P[q \leq t] = 1 - \rho e^{-t/W}, \qquad P[q > t] = \rho e^{-t/W}.$$

$$W_q = E[q] = \rho E[s]/(1 - \rho).$$

$$\sigma_q^2 = (2 - \rho)\rho E[s]^2/(1 - \rho)^2.$$

$$E[q | q > 0] = W, \qquad \mathrm{Var}[q | q > 0] = W^2.$$

$$\pi_q(90) = W \ln(10\rho), \qquad \pi_q(95) = W \ln(20\rho).$$

$$\pi_q(r) = W \ln\left(\frac{100\rho}{100 - r}\right).$$

All percentile formulas for q will yield negative values when ρ is small: all negative values should be replaced by zero. For example, if ρ is 0.02, then 98 percent of all customers do not have to queue for service so the 95th percentile value of q is zero: so are the 90th and 95th percentile values.

TABLE 4

Steady State Formulas for M/M/1/*K* Queueing System

($K \geq 1$ and $N \leq K$)

$$p_n = P[N = n] = \begin{cases} \dfrac{(1-u)u^n}{1-u^{K+1}} & \text{if } \lambda \neq \mu \text{ and } n = 0, 1, \ldots, K \\ \dfrac{1}{K+1} & \text{if } \lambda = \mu \text{ and } n = 0, 1, \ldots, K. \end{cases}$$

$p_K = P[N = K]$. Probability an arriving customer is lost.

$\lambda_a = (1 - p_K)\lambda$. λ_a is the actual arrival rate at which customers enter the system.

$$L = E[N] = \begin{cases} \dfrac{u[1-(K+1)u^K + Ku^{K+1}]}{(1-u)(1-u^{K+1})} & \text{if } \lambda \neq \mu \\ \dfrac{K}{2} & \text{if } \lambda = \mu. \end{cases}$$

$L_q = E[N_q] = L - (1 - p_0)$

$$q_n = \frac{p_n}{1 - p_K}, \qquad n = 0, 1, 2, \ldots, K-1.$$

q_n is the probability that there are n customers in the system just before a customer enters.

$$W(t) = P[w \leq t] = 1 - \sum_{n=0}^{K-1} q_n \sum_{k=0}^{n} e^{-\mu t} \frac{(\mu t)^k}{k!}.$$

$W = E[w] = L/\lambda_a$.

$$W_q(t) = P[q \leq t] = 1 - \sum_{n=0}^{K-2} q_{n+1} \sum_{k=0}^{n} e^{-\mu t} \frac{(\mu t)^k}{k!}.$$

$W_q = E[q] = L_q/\lambda_a$.

$E[q \mid q > 0] = W_q/(1 - p_0)$.

$\rho = (1 - p_K)u$.

ρ is the true server utilization (fraction of time the server is busy).

$u = \lambda/\mu = \lambda E[s]$, $\quad \rho = u/c$.

$$p_0 = P[N = 0] = \left[\sum_{n=0}^{c-1} \frac{u^n}{n!} + \frac{u^c}{c!\,(1-\rho)}\right]^{-1} = c!\,(1-\rho)C(c, u)/u^c.$$

$$p_n = \begin{cases} \dfrac{u^n}{n!} p_0 & \text{if } n = 0, 1, \ldots, c \\[2ex] \dfrac{u^n p_0}{c!\, c^{n-c}} & \text{if } n \geq c. \end{cases}$$

$$L_q = E[N_q] = \lambda W_q = \frac{uC(c, u)}{c(1-\rho)}, \qquad \sigma^2_{N_q} = \frac{\rho C(c, u)[1 + \rho - \rho C(c, u)]}{(1-\rho)^2}.$$

where $C(c, u) = P[N \geq c]$ = probability all c servers are busy is called Erlang's C formula.

$$C(c, u) = \frac{u^c}{c!} \Bigg/ \left[\frac{u^c}{c!} + (1-\rho)\sum_{n=0}^{c-1} \frac{u^n}{n!}\right].$$

$$L = E[N] = L_q + u = \lambda W.$$

$$W_q(0) = P[q = 0] = 1 - \frac{p_c}{1-\rho} = 1 - C(c, u).$$

$$W_q(t) = P[q \leq t] = 1 - \frac{p_c}{1-\rho} e^{-ct(1-\rho)/E[s]} = 1 - C(c, u)e^{-t/E[q|q>0]}.$$

$$W_q = E[q] = \frac{C(c, u)E[s]}{c(1-\rho)}, \qquad E[q \mid q > 0] = \frac{E[s]}{c(1-\rho)}.$$

$$\sigma_q^2 = \frac{[2 - C(c, u)]C(c, u)E[s]^2}{c^2(1-\rho)^2}. \qquad \pi_q(r) = \frac{E[s]}{c(1-\rho)} \ln\left(\frac{100C(c, u)}{100 - r}\right).^*$$

$$\pi_q(90) = \frac{E[s]}{c(1-\rho)} \ln(10C(c, u)), \qquad \pi_q(95) = \frac{E[s]}{c(1-\rho)} \ln(20C(c, u)).^*$$

$$W(t) = P[w \leq t] = \begin{cases} 1 + C_1 e^{-\mu t} + C_2 e^{-c\mu t(1-\rho)} & \text{if } u \neq c - 1 \\ 1 - [1 + C(c, u)\mu t]e^{-\mu t} & \text{if } u = c - 1. \end{cases}$$

where $C_1 = \dfrac{u - c + W_q(0)}{c - 1 - u}$ and $C_2 = \dfrac{C(c, u)}{c - 1 - u}$.

$$W = E[q] + E[s].$$

$$E[w^2] = \begin{cases} \dfrac{2C(c, u)E[s]^2}{u + 1 - c}\left\{\dfrac{1 - c^2(1-\rho)^2}{c^2(1-\rho)^2}\right\} + 2E[s]^2. & u \neq c - 1. \\[2ex] 4C(c, u)E[s]^2 + 2E[s]^2. & u = c - 1. \end{cases}$$

$$\sigma_w^2 = E[w^2] - E[w]^2$$

$$\left.\begin{aligned} \pi_w(90) &\approx W + 1.3\sigma_w \\ \pi_w(95) &\approx W + 2\sigma_w \end{aligned}\right\} \text{Martin's estimates}$$

* All percentile formulas for q yield negative values for low server utilization: all should be replaced by zero.

Steady State Formulas for M/M/2 Queueing System

$\rho = \lambda E[s]/2 = u/2.$

$p_0 = P[N = 0] = (1 - \rho)/(1 + \rho).$

$$p_n = P[N = n] = 2p_0\rho^n = \frac{2(1-\rho)\rho^n}{(1+\rho)}, \qquad n = 1, 2, 3, \ldots.$$

$$\left.\begin{aligned} \pi_w(90) &\approx W + 1.3\sigma_w \\ \pi_w(95) &\approx W + 2\sigma_w \end{aligned}\right\} \text{ Martin's estimate.}$$

$$L_q = E[N_q] = \frac{2\rho^3}{1-\rho^2}, \qquad \sigma^2_{N_q} = \frac{2\rho^3[(\rho+1)^2 - 2\rho^3]}{(1-\rho^2)^2}$$

$C(2, u) = P[\text{both servers busy}] = 2\rho^2/(1 + \rho).$

$L = E[N] = L_q + u = 2\rho/(1 - \rho^2).$

$W_q(0) = P[q = 0] = (1 + \rho - 2\rho^2)/(1 + \rho).$

$W_q(t) = P[q \le t] = 1 - [(2\rho^2)/(1 + \rho)]e^{-2\mu t(1-\rho)}.$

$W_q = E[q] = \rho^2 E[s]/(1 - \rho^2), \qquad E[q \mid q > 0] = E[s]/2(1 - \rho).$

$\sigma_q{}^2 = \rho^2(1 + \rho - \rho^2)E[s]^2/(1 - \rho^2)^2.$

$$\pi_q(r) = \frac{E[s]}{2(1-\rho)} \ln\left(\frac{200\rho^2}{(100 - r)(1 + \rho)}\right).$$

$$\pi_q(90) = \frac{E[s]}{2(1-\rho)} \ln\left(\frac{20\rho^2}{1+\rho}\right), \qquad \pi_q(95) = \frac{E[s]}{2(1-\rho)} \ln\left(\frac{40\rho^2}{1+\rho}\right).^a$$

$$W(t) = P[w \le t] = \begin{cases} 1 - \dfrac{(1-\rho)}{1 - \rho - 2\rho^2} e^{-\mu t} + \dfrac{2\rho^2}{1 - \rho - 2\rho^2} e^{-2\mu t(1-\rho)} & \text{if } u \neq 1 \\ 1 - \left[1 + \dfrac{\mu t}{3}\right] e^{-\mu t} & \text{if } u = 1. \end{cases}$$

$W = E[s]/(1 - \rho^2).$

$$E[w^2] = \begin{cases} \dfrac{\rho^2 E[s]^2[1 - 4(1-\rho)^2]}{(2\rho - 1)(1 - \rho)(1 - \rho^2)} + 2E[s]^2, & u \neq 1. \\ \frac{10}{3}E[s]^2, & u = 1. \end{cases}$$

$\sigma_w{}^2 = E[w^2] - E[w]^2.$

[a] All percentile formulas for q yield negative values for low server utilization; all such should be replaced by zero.

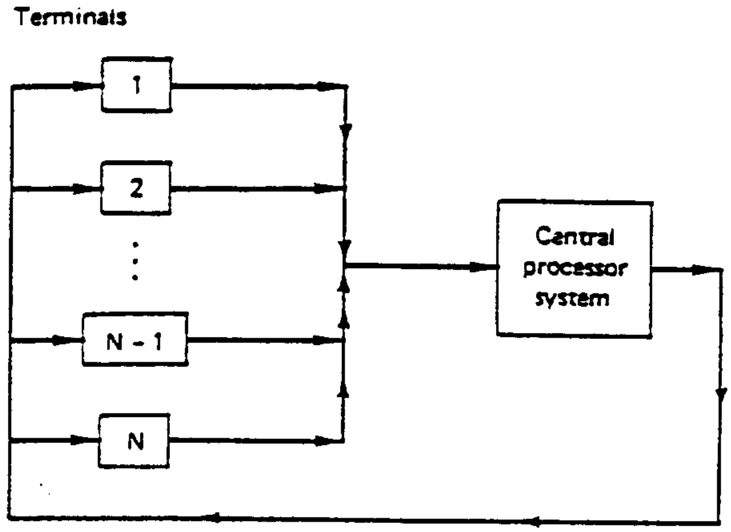

Fig. 6.3.1 Finite population queueing model of interactive computer system. Special case in which the central processor system consists of a single CPU with processor-sharing queue discipline.

The CPU operates with the processor-sharing queue discipline. CPU service time is general with the restriction that the Laplace–Stieltjes transform must be rational. The same restriction holds on think time. $E[t] = 1/\alpha$ is the average think time with $E[s] = 1/\mu$ the average CPU service time. Then

$$p_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{E[s]}{E[t]}\right)^n\right]^{-1} = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\alpha}{\mu}\right)^n\right]^{-1}.$$

The CPU utilization

$$\rho = 1 - p_0,$$

and the average throughput

$$\lambda_T = \frac{\rho}{E[s]} = \frac{1-p_0}{E[s]}.$$

The average response time

$$W = \frac{NE[s]}{1-p_0} - E[t].$$

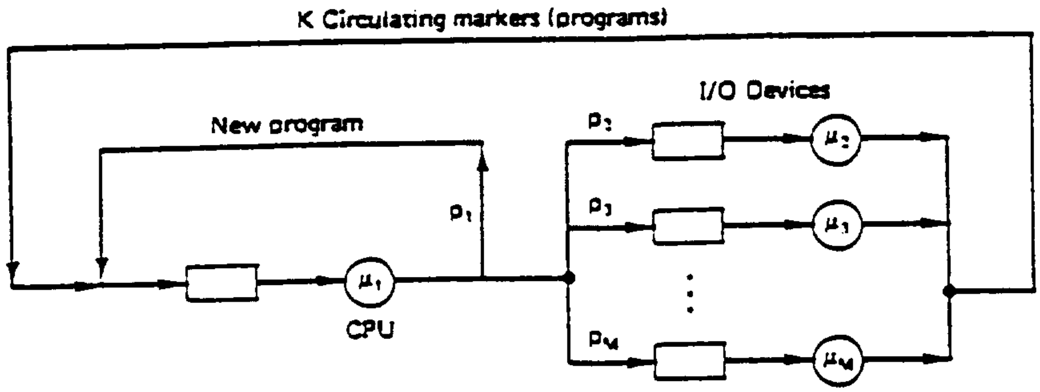

Fig. 6.3.4 Central server model of multiprogramming.

For the assumptions of the model see Section 6.3.4.
Calculate $G(0), G(1), \ldots, G(K)$ by Algorithm 6.3.1 (Buzen's Algorithm).
Then the server utilizations are given by

$$\rho_i = \begin{cases} G(K-1)/G(K) & i = 1 \\ \dfrac{\mu_1 \rho_1 p_i}{\mu_i} & i = 2, 3, \ldots, M. \end{cases} \tag{6.3.25}$$

The average throughput λ_T is given by

$$\lambda_T = \mu_1 \rho_1 p_1. \tag{6.3.26}$$

If the central server model is the central processor model for the interactive computing system of Fig. 6.3.1. then the average response time W is calculated by

$$W = \frac{N}{\lambda_T} - E[t] = \frac{N}{\mu_1 \rho_1 p_1} - E[t]. \tag{6.3.27}$$