Network Model of the Processor System

VLADIMIR HOTTMAR

The University of Zilina, Department of Telecommunications, Slovak Republic
hottmar@fel.utc.sk

Keywords: processor, queuing theory, service centre, closed network, dual port memory, data request

The paper deals with modelling of a double-processor system by closed service network. The article presents the queuing system as a method of modelling. It analyses the performance and quantifies time characteristics of the processor systems.

1. Introduction

Questions connected with monitoring and valuation of processor systems (PS) are actual in many different levels of it's utilization. One of many possible processor system applications appears from assumption of dividing the final number of processed tasks among fractional processor systems, which activity is mutually and relatively independent, and every processor system Ps_j of general system M process certain class of tasks r_i of total set R, where $Ps_j \in M$ and $r_i \in R$, when i,j = 1, 2,...

Our task will be to understand a certain input set of data processed by given programme, and activities resulting from this processing will be the output set of data. Data transmission among processor systems is realised in the form of messages S that, on basis of routing and information content of message, is possible to consider as requests (answers) N. Requests are in every processor system saved into memory where they create a family of requirements. Systems processors solve the operation of requests. Lets define data transmission (independently of its information content) as requests coming (outgoing) into (from) PS and lets define processing of messages as operation of requests by processor system. Assign to the set of data the set of requests $S \rightarrow N$, where

S = {1, 2,...s} a N{1, 2,...p} and to the subset of processors (P_i) the set of service systems – service centres Σ_i. Then $\{P_i\} \rightarrow \{\Sigma_i\}$, where i = 1, 2,...

Computing memories are specific. Function of these memories is to save results and intermediate results of operations with data. As the processor always has free access to all data in the memory, it is needful to define the strategy of saving and selection of requests into (from) the memory. Then it is possible to transform operation memories to the family of requests, and to define the strategy of its service. Service strategies are worked out in detail in [1,2,3,4].

Assignment of the set of data to the set of requests, and the set of processors to the set of service centers allow to model every local processor system by stochastic model and to solve this model by queuing theory.

Lets apply queuing theory to a simple structure of a double-processor system with processors P_1 and P_2. Block structure of double-processor system and its model are shown in *Figure 1*. Processors perform their activity autonomously and the reciprocation of data through dual port memory. Reciprocation of data with the environment they can perform independently from each other. Data entry from environment (terminals) is realised only through input processor circuit P_1.

Processors P_1 and P_2 process tasks, and transmission of tasks between service centres are defined by probabilities of transition *pji*.

Terminals T_1, T_2,..., T_n generate demand for service, which are serially stored in queue L_1. As the system doesn't distinguish between priority or non-priority requests, the requests for service will be selected from queues L_1, L_2 serially, as they are arriving. L_1, L_2 present the set of requests with FIFO service strategy. Service centres Σ_1, Σ_2 perform service of requests. If the request is serviced in service centre Σ_1 then it leaves from the service system with probability p_{11}, or with probability p_{12} it requires the service from service centre Σ_2. If Σ_2 is free, the requests are served immediately, otherwise it takes stand in the queue L_2. After servicing has finished, the requests leave the service system with probability p_{2T}, or with probability p_{12} they come to the queue L_1 and they again require the service from service centre Σ_1. After the service has finished in Σ_1, they leave from system with p_{21} probability, or with p_{12} probability they require another service from the service centre Σ_2, etc...

The system and its network model are shown in *Figure 1*.

2. Network model of double-processor system

Network model comes from the principle of **closed service network**. Characteristic feature of closed service network is the uniformity of total number of requests. According to this, the intensities of arrivals into particular service centres are non-constant.

The state of closed service network is at every instant of time defined as

$$K = \sum_{i=1}^{N} k_i \tag{1}$$

where
N – is a total number of network nodes (Figure 1),
K – is a total number of requests in network,
k_i – is number of requests in i-th node.

We define the total number of network states as

$$Z = \binom{N+K-1}{N-1} \tag{2}$$

Local balance condition [1],[2] is the theoretical presumption for existence of closed service network solution by service centre. If every service centre in network fulfills the local balance condition, then even the whole network fulfil the balance condition. A solution is possible for service networks composed of local service centres of M/M/n/FIFO type [1],[2]. A precise solution of closed service networks was given by Gordon and G. F. Newell [1] and is defined as

$$p(k_1, k_2, \ldots k_N) = \frac{1}{G(K)} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta(k_i)} \tag{3}$$

which is the probability that k_i requests are in i-th service centre, where x_i can be obtained from equation system solution

$$\mu_i x_i = \sum_{j=1}^{N} \mu_j x_j p_{ji} \tag{4}$$

where i = 1, 2,...,N and $G(K)$
is a **normalization constant**.

If every service centre has only one service device, then $\beta(k_i)$ = 1 and (3) changes to the form

$$p(k_1, k_2, \ldots k_N) = \frac{1}{G(K)} \prod_{i=1}^{N} x_i^{k_i} \tag{5}$$

while, for the stationary state probability vectors, the following condition has to be fulfilled:

$$\sum_{S_K} p(k_1, k_2, \ldots k_N) = 1 \tag{6}$$

Closed service network is characterized by input parameters:
N – number of service centres (nodes) of model service network
K – total number of requests in the model
μ_i – service intensity of requests processing in i-th service centre, i = 1, 2,...,N
p_{ji} – transmission probability, that means – the request for service in j-th service centre will also require service in i-th service centre.

If we put j = 1 and i = 2,3,...,N in equation system (4) then this will take the form

$$x_i \mu_i = \mu_1 x_1 p_{1i}, \tag{7}$$

or

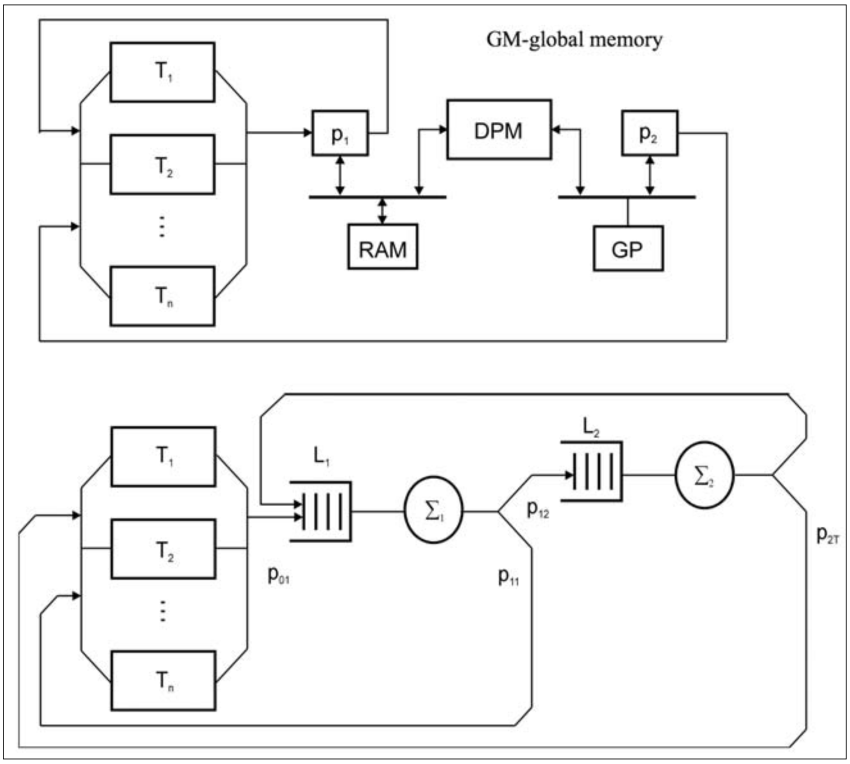
$$x_i = \frac{\mu_1}{\mu_i} x_1 p_{1i} \quad \text{for } i = 2, 3, \ldots, N \tag{8}$$

For x_i we get an equation system of (N-1) equations which are independent. This equation system enables to express unknown x_i through the parameter, for instance x_1.

If x_1=1, the normalization constant we define from equation

$$G(K) = \sum_{i=1}^{N} \prod_{i=1}^{N} x_i^{k_i} \tag{9}$$

Figure 1. Block structure of computing system and its network model



The basic indicator is utilization (loading) ρ_1, ρ_2 of service centres Σ_1 and Σ_2 [1,3,7]. By analysing (9) we arrive to the following conclusion. Exponents k_i define possible number of requests in i-th service centre. Marginal attributes for k_i are zero (condition, when i-th service centre is empty) and K (all service centres are empty accept i-th). Consequently, two following alternatives can occur for k_i

$$(k_i = 0) \cup (k_i = K) \tag{10}$$

If we want to define ρ_1 then we consider only those states, in which k_1>0, that means, k_1=K for i = 2,3,...,N can never happen at the some time.

As the minimum of requests in first service centre (i =1) is k_1=1, then the maximum

$$k_{i_{\max}} = (K-1) \tag{11}$$

for i = 2,3,...,N is divided into the rest of service centres [2],[9].

According to (9) we can write

$$\sum_{S_K}\prod_{i=2}^{N} x_i^{k_i} = G\left(k_{i_{\max}}\right) = G(K-1) \qquad (12)$$

Utilization of i-th service centre is defined as

$$\rho_i = \sum_{S_K:k_i\rangle 0} p\left(k_1,k_2,...k_N\right) = \sum_{S_K:k_i\rangle 0} \frac{1}{G(K)}\prod_{i=2}^{N} x_i^{k_i} \qquad (13)$$

Then the utilization of ρ_1 service centre $i=1$ in terms of (12) and (13)

$$\rho_1 = \sum_{S_K:k_1\rangle 0} p\left(k_1,k_2,...k_N\right) = \sum_{S_K:k_1\rangle 0} \frac{1}{G(K)}\prod_{i=2}^{N} x_i^{k_i} = \frac{G(K-1)}{G(K)} \qquad (14)$$

We can interpret the meaning of (14) as follows.

Utilization of ρ_1 service centre $i=1$ is given by summary of state probabilities in which k_1 values larger than zero $k_1>0$ are gained. As an event, for $k_1=0$, surely occurs once at least, for ρ_1 we can write as

$$\rho_1 = \frac{G(K-1)}{G(K)} \qquad (15)$$

Utilization of ρ_1 service centres $i=2,3,...,N$ we get from balance condition of input and output flow intensities of requests in steady-state regime in i-th service centre. Input flow intensity from service centre i is given by summary of service intensity μ_I and probability, that i-th service centre is occupied, which is just ρ_i.

Concerning (8) and (15) the following equation is valid

$$\rho_i = x_i.\rho_1 \quad or \quad \rho_i = x_i\frac{G(K-1)}{G(K)} \qquad (16)$$

Intensities of λ_i requests arriving in individual service centres are defined for steady state of input and output request flow equation of local service centre. For λ_I will be valid:

$$\begin{aligned}\lambda_i &= \rho_i\mu_1 \quad for\ i=1\\ \lambda_i &= \rho_1\mu_1 p_{1i} \quad for\ i=2,3,...,N\end{aligned} \qquad (17)$$

Average amount of requests K_i in i-th service centre will be defined from the condition

$$K_i = \sum_{k_{ie}\rangle 0} k_{ie}\, p\left(k_1,k_2,...k_N\right) \qquad \text{for} \quad e=1,2,\ldots,K \qquad (18)$$

where k_{ie} is number of requests in e-th service centre and $p(k_1,k_2,...,k_N)$ is the already known vector of state line probabilities. We perform the summation through lines, in which $k_{ie}>0$, and according to (1), condition of requests uniformity in closed network has to be fulfilled. We define the average length of L_i line in i-th service centre from deference of serviced ρ_I and from average number of requests waiting for service, provided that number of service devices (processors) in i-th service centre is equal to one.

$$L_i = K_i - \rho_i \qquad \text{for} \quad i=1,2,\ldots,N \qquad (19)$$

Time characteristics of service centres – the average stopping time of T_i requests in i-th service centre and average waiting period of T_{wi} requests in L_i line we can define from Little's law:

$$T_i = \frac{K_i}{\lambda_i} \qquad (20)$$

and

$$T_{wi} = \frac{L_i}{\lambda_i} = T_i - \frac{1}{\mu_i} \qquad (21)$$

3. Model analysis

We analyse the model of *Figure 1*. We assume a real technical environment and we set a number of connected terminals to 9, as an example (T_1 till T_9). Implementation of μ_i and p_{ji} parameters will help us to define all performance parameters and time characteristics of the system. The number of closed network states for K=9 and N=2 define (2), according to which Z=10.

From system (4) we define the equations for parameters x_i calculation, for two service centres, then

$$\begin{aligned} x_2\mu_2\left(p_{21}+p_{2T}\right) &= x_1\mu_1 p_{12}\\ x_1\mu_1\left(p_{12}+p_{11}\right) &= x_2\mu_2\left(p_{2T}+p_{21}\right) + x_1\mu_1 p_{11}\end{aligned} \qquad (22)$$

for complete probabilities it is valid that

$$p_{12}+p_{11}=1, \quad p_{21}+p_{2T}=1 \qquad (23)$$

By elimination of equations for $x_1=1$ we get

$$x_2 = \frac{\mu_1}{\mu_2}p_{12} \qquad (24)$$

We see that parameter x_2 is dependent on ratio μ_1/μ_2 and on p_{12} probability. These three variables crucially affect performance parameters of the system model. Let's analyse the model when the service intensity μ_1,μ_2 of both processors P_1, P_2 are the same, and requests for service in service centre Σ_1 demand the service in service centre Σ_2 with p_{12}=1/2 probability. According to (24), the parameter x_2 will obtain value x_2=0,5.

Substitution of x_1 and x_2 into (9) is normalization constant

$$G(K) = G(9) \doteq 1{,}998046$$

For performance parameters definition we use numerical figures from *Table 1*. Table shows in the first column the number of S network states, in second and third column we can see the distribution of requests k_1 and k_2 in service centres i = 1, 2.

Table 1.

S	k_1	k_2	$x_1^{k_1}x_2^{k_2}$	$p(k_1,k_2)$
1	0	9	0,001953	0,000977
2	1	8	0,003906	0,001955
3	2	7	0,007812	0,00391
4	3	6	0,015625	0,00782
5	4	5	0,03125	0,01564
6	5	4	0,0625	0,03128
7	6	3	0,125	0,062561
8	7	2	0,25	0,125122
9	8	1	0,5	0,250244
10	9	0	1,0	0,500488
Σ			1,998046	0,999997

The fourth column defines on the right side of the relation fractional conjunction of (9) and in the last column there are values of state line probabilities.

The value $G(K\text{-}1)$ is defined for $i=1$ from (12), or from Table 1 for $G(8)$=1,99414, then from (15) we define $\rho_1 \doteq$ 0,998, and from (16) we get by substitution of ρ_1 the value for $\rho_2 \doteq$ 0,499.

For the definition of λ_1, λ_2 we need to know the numerical value for service intensity μ_1. Then by means of (17) the average number of requests in service centres will be K_1=8,00975 and K_2=0,990214, while the relation (1) has to be valid for total number of requests in network, that means K=8,999964.

The average length of lines L_1=7,01175 and L_2= 0,491214 requests waiting for service. We will define time characteristics of T_1, T_2 and T_{W1}, T_{W2} from well known λ_1 and λ_2, and from (20) and (21).

Now let's analyze the model in the case when the service intensities are mutually deferent $\mu_1 \neq \mu_2$, and for probability of transition it will be again valid p_{12}=1/2. On the basis of (24) x_2=1, and x_1=1 we leave unchanged. After replacement x_1 and x_2 into (5) we obtain numerical data shown in *Table 2*.

Table 2.

S	k_1	k_2	$x_1^{k_1} x_2^{k_2}$	$p(k_1,k_2)$
1	0	9	1	0,1
2	1	8	1	0,1
3	2	7	1	0,1
4	3	6	1	0,1
5	4	5	1	0,1
6	5	4	1	0,1
7	6	3	1	0,1
8	7	2	1	0,1
9	8	1	1	0,1
10	9	0	1	0,1
Σ			10	1

Achievement parameters:

$$G(9) = 10,\ G(8) = 9,\ \rho_1 = \rho_2 = 0{,}9,\ K_1 = K_2 = 4{,}5,\ K = K = 9,\ L_1 = L_2 = 3{,}6,\ \lambda_1 = 0{,}9\ \mu_1,\ \lambda_2 = 0{,}45\ \mu_1$$

4. Discussion of the results

Our conclusions are as follows.

- Equal service intensities ($\mu_1 = \mu_2$) cause different coefficients of utilization of ρ_1 and ρ_2, while service centre Σ_1 (that is the processor P_1) works on saturation limit and creates a bottleneck in the system [7], the processor P_2 works with approximately 50% utilization.
- Coefficient $\rho \to 1$, which means that we can expect a breach of local balance in model (the steady state stops to be valid), what will be at the real processor P_1 expressed by not accepting the request for service and by memory conflicts.
- Unequal dividing of the requests in lines L_1, L_2 represent the increase of memory claims at real memories.
- Change in service intensity ratio (μ_1/μ_2=2) at p_{12}= 1/2 result in equal dividing of requests K_1,K_2 in service centres Σ_1 and Σ_2, utilization ratio ρ_1, ρ_2 gain equal values, and the processors P_1 and P_2 work with approximately 90% utilization.

Acknowledgement

The author gratefully acknowledges the support by the VEGA project No. 1/2043/05 "Systems for the interactive television cabling distribution ICATV"

References

[1] Rukovansky, I.:
Valuation of computer system efficiency, 1989., SNTL, pp.61–97.

[2] Neuschl, S. et al.:
Modelling and simulation, pp.304–354.

[3] Mitrani, I.:
Modelling of computer and communication systems, Cambridge University Press, 1987., pp.73–77.

[4] Zitek, F.:
Lost time, Academia Prague 1969, pp. 42–47.

[5] Kluvanek, P. and Brandalik, F.:
Operation analyse I. Bulk service theory, pp.40–42.

[6] Pesko, S. and Smiesko, J.:
Stochastic models of operation analyze, University of Zilina, 1999., pp.75–80.

[7] Hottmar, V.:
Performance model of network calculating systems ITKR, University of Zilina / EDIS – editorship ZU, 2005. ISBN 80-8070-338-8

[8] Hottmar, V.:
Digital television – text-book from international seminar, University of Zilina, May 1999., pp.66–90.

[9] Robertazzi, T. G.:
Computer Networks and Systems: Queuing Theory and performance Evaluation, Springer-Verlag New York, 1990, pp.181–233.