MTA Számítástechnikai és Automatizálási Kutató Intézete, Közlemények 27/1982

THROUGHPUT OPTIMIZATION OF MULTISTAGE, QUEUEING SYSTEMS WITH FINITE INTERMEDIATE STORAGE*)

ADAM WOLISZ+)

1. Introduction .

Multistage Queueing Systems /MQS/, that is such systems where every demand has to be served consecutively, in a predefined order, by several servers receive recently a lot of attention, being an important tool for modelling several kinds of industrial systems. For example such structure have production lines [6] and computer communication systems [12] .
If service times at different stages are not constant and equal then unavoidably some queues form between consecutive servers. In reality this queues are not allowed to exceed some fixed values because of storage facility constraints. If, upon completion of a service, no place is available in a consecutive buffer for depositing the demand involved, this very demand may perhaps be lost, and abandon the system never to return again.

*) The paper has been presented at VI th Conference on Operating Systems, hold at Visegrad, Hungary, in February 1980.

+) Polish Academy of Sciences, Department of Complex Automation Systems
44-100 Gliwice, Zwycięstwa 21, Poland.

More frequently however, no losses are permitted and server which completed the "fatal" service is used as an additional storage place being of course unable to process other demands /the blocking phenomenon/.

In fact different possible types of blocking may occure, like repeating the service of demand which couldn't have been placed in the consecutive buffer, or even forcing this demand to return to the very beginning of the system and have all the so far achieved servic repeated / cf [11] where also some equivalence rules between differe types of blocking where discussed and [31] /.

Numerous papers were concerned with the analysis of MQS.It has to be stressed that exact analytical tools generally fail when the number of consecutive servers exceeds three, and even for smaller systems only some special cases / or some special system features/ are fully investigated.Thus a great affort is being done to obtain approximate solutions either by means of specially developed methods like diffusion approximation /cf [24] /, numerical methods /eg [19] or simulation.

Other, equally important area of research is optimization of MQS operational features, like throughput, servers utilization, queue length e.t.c.For these studies usually the following way was chosen by numerous researchers: First the special cases of two and three stage systems were investigated /preferably analytically/ and afterwards, basing on conclusions obtained there some hypothesis of more general applicability were formulated.These in turn were subjec to verification using- most frequently - simulation as the tool.

The purpose of this paper is to present state of the art in the area of throughput optimization in MQS with the classical type of blocking mentioned above.In author's opinion there is a need for such survey as in a number of papers several optimization rules have been developed /usually each for some special case/, being sometimes non-consistent or even contradictory.

In consecutive sections, after dealing with some general properties including the formal statement of the optimization problem and with the special case of systems having constant service times, outlines for optimal choice of each of the parameters influencing the throughput of MQS with single servers at every stage will be considered in turn. This will be followed by considerations concerning the use of multiservers at some stages, and some remarks about optimization goals other than throughput maximization. The whole paper is completed by a set of final conclusions.

The list of references compiled in this paper, although not aimed as a complete bibligraphy includes, in author's opinion, the vast majority of papers concerned with optimization problems in unpaced MQS /i.e. such where no external synchronization in operation of different stages exists/. Papers dealing with paced systems were mentioned only if the results presented there were in strong connection with the investigated system, while papers covering the problems of system analysis exclusively, have been intentionally omitted.

As production lines are one of the common technical systems being modelled as MQS, it is worth mentioning, that a wide range of both analysis and optimization problems connected with designing of production lines was reviewed by Buxey, Slack and Wild [6] establishing also their connection with topics discussed here.

It is hoped that the unified approach presented here will be of some help in directing the future research, simultaneously providing practicians with a set of directly applicable optimization rules.

2. Concepts and Definitions .

Further in this paper MQS of the type presented in Fig. 1 will be considered.

Identical demands originating from a source $\mathcal{W}$ with intensity λ_o are to be served consecutively on "M" stages /each of them consisting of several, not necesserily similar service facilities/ in a strict order.A queue S_i with N_i places is allowed to build up in front of the i-th service stage.The service time of demands on server A_i^j are independent,identically distributed nonnegative random variables b_i^j, with arbitrary distribution functions $B_i^j(x)$.

Let $E\ (b_i^j)$ denote the mean service time and μ_i^j its reciprocal /service intensity/, $\mu_i^j = \left[E\ (b_i^j)\right]^{-1}$ .Random variables b_i^j and b_l^k are statistically independent if $j \neq k$ or $i \neq l$.

It is assumed, that only one demand can be served by a server at any time.

Each stage is preceded by a buffer.Intermediate buffers S_2 , ... , S_M are of finite size $/N_i < \infty$, $i = 2, 3, \ldots M/$, causing the blocking phenomenon to occure.

Each server A_i^j , $j = 1, 2, \ldots n_i$; $i = 1, 2, \ldots , M-1$ is always in one of three possible stages:

- busy if it is serving a demand;
- blocked, when it has completed a service but cannot pass on the demand to the next stage, because the consecutive buffer S_{i+1} is full;
- idle when it is neither busy nor blocked.

We shall assume that the server A_i^j may be idle if and only if there are no demands waiting for service in the queue S_i.

Let us now introduce some classification of MQS.

A MQS will be called queueing line if every stage consists of one server, exclusively $/n_i = 1 ; i = 1,2, \dots , M/$. Then $B_i^1(x)$ will be denoted briefly $B_i(x)$.

If all servers installed at any stage are identical $/ B_i^j(x) = B_i^*(x) ; j = 1, 2, \dots , n_i ; i = 1, 2, \dots M /$ then the MQS is called homogeneous, otherwise it will be referred to as non-homogeneous .

If the joint service intensity of all servers installed at stage "i" , i= 1, 2, ... , M is constant,

$$\sum_{j=1}^{n_i} \mu_i^j = D , \qquad i = 1,2, \dots M; \qquad (1)$$

then the system is called balanced, otherwise it is unbalanced .

In order to preserve a measure of system unbalancing, we shall further assume that the following holds:

$$\sum_{i=1}^{M} \left[\left(\sum_{j=1}^{n_i} \mu_i^j \right) \right]^{-1} = M \qquad (2)$$

Thus for the balanced case

$$\sum_{j=1}^{n_i} \mu_i^j = 1 . \qquad (3)$$

Naturally for homogeneaus, balanced systems ;

$$\begin{aligned} \mu_i^j &= \mu_i^* & j &= 1,2,\dots, n_i \\ \mu_i^* \cdot n_i &= 1 & i &= 1,2, \dots, M \end{aligned} \qquad (4)$$

Comparing the service time distribution effect on the system throughput we shall frequently utilize the variability coefficient, defined for a service time distribution B (x) as

$$C^2 = \frac{\int_0^\infty x^2 dB(x)}{\left[\int_0^\infty x dB(x) \right]^2} - 1 , \qquad (5)$$

and being a suitable measure for the variability of service time. Naturally $\zeta=0$ for the constant service time and $\zeta = 1$ for exponentially distributed service time.

In some queueing lines the transfer of demands from all servers to consecutive buffers takes place simultaneously, being externally synchronized, no matter if all service processes where completed or not.This occurs for example in automated transfer / moving belt/ production lines.Such queueing lines will be called paced .

Thus in the paced systems a predetermined time quantum, /called cycle/ is imposed for every service.In the case of constant service times the cycle should be equal to the longest service time. If service times are variable then the line should be designed so as to minimize the probability of one or more stations exceeding the cycle.

Methods for designing paced queueing lines were surveyed in [6], while the case of variable service times is treated for example in [43] .

Further in this paper we shall be interested only in the cases when no external synchronization / no pacing effect/ in the flow of demands through the system is introduced, called unpaced systems.

An important feature of any queueing system is its throughput /frequently called production rate/ defined as the mean number of demands leaving the system in a time unit /completly served/.

We shall be specially interested in the maximal throughput / capacity/ which a given system may achieve.

The MQS'is called open if the buffer S_1 preceding the first service stage is unlimited, $/N_1 = \infty /$, and the input stream of demanc is a renewal stream.

On the other hand if the input stream is such, that the queue S_1 is never empty, then the system is called saturated. Certainly for saturated systems both the maximal queue size N_1 and the detailed characteristics of the input stream are of no importance.

Two MQS : an open one and a saturated one are called corresponding if they are identical, up to the specification of the input stream.

Let us consider some MQS. If we define another system in which the order of service is reversed, that means every demand passes through the system beginning with stage M and ending at stage 1, buffer sizes being exactly preserved, then such two systems are called dual systems .

Finally we shall introduce the concept of saturated system accumulation , being equal to the maximal number of demands which are allowed in the system simultaneously.
System accumulation V is given by the following formula

$$L = \sum_{i=1}^{M} n_i + Z , \qquad (6)$$

where

$$Z = \sum_{i=2}^{M} N_i$$

is the total number of storage places /total buffer size/ available in the system.

It is evident that dual systems have equal accumulation.

3. General considerations .

In this section we shall present some theorems and remarks concerning some large classes of MQS.

Lavenberg [28] proved / as a special case of more general dependences discussed in his paper/ an important connection between the throughputs of corresponding open and saturated multistage syste If only all service stations in a MQS have Coxian service time distr butions*) and all intermediate buffers are finite, then throughput of the saturated system T_s is equal to the maximal throughput /capacity/ V of a corresponding open system.

More precisely, the throughput T of the open system can be defined as follows:

$$T = \begin{cases} \lambda_0 & \text{if} \quad \lambda_0 < T_s \\ T_s & \text{if} \quad \lambda_0 > T_s \end{cases}$$

where λ_0 denotes /cf Fig.1/ the intensity of demands arrival in the open system. Thus $V = T_s$.

Additionally in the case when $\lambda > T_s$, the stationary distribution o the number of demands waiting in the queue S_1 doesn't exist, contr ry to the case when $\lambda < T_s$.

Thus in order to find the capacity of multistage queueing syste it is necessary to investigate the proper saturated case.

*) Coxian distribution means any distribution having a rational Laplace transform of the distribution function. In fact it is possible to approximate any distribution function fairly well with a Coxian distribution, thus the constraints are not restrictive. The case of constant service times, however also possible to approach as a limiting case of Erlangian distribution, will be further in this paper traated seperately.

Let us notice that without loss of generality it is enough to consider MQS without intermediate buffers. ([2], [31]).
It is quite evident, that any buffer of size N can be replaced by N servers, each having a null service time / $B(x) = \mathbb{1}(x)$, $\mathbb{1}(x)$ being the Heaviside function/ .This property simplifies many proofs.

It is also worth stressing, that the capacity of a MQS is almost independent of the service disciplines applied.In fact, as long as no server A_i^j , $j = 1, 2, \dots , n_i$; $i = 1, 2, \dots M$ may remain idle if the queue S_i is not empty, and the service is of nonpreemptive type,all queueing disciplines / possibly not identical at different stages/ yield equal system capacity.

Another important result of quite general applicability is the so called reversibility property for saturated MQS with finite intermediate buffers.
Yamazaki and Sakasegawa [54] demonstrated that the capacity of a queueing line with general service time distribution is invariant for reversal ordering of the servers.That means, dual systems have equal capacity.

The reversibility property is important for the throughput optimization considerations as it makes us expect optimization rules calling for queueing line structures being in some sense "symmetric". This feature will become more meaningfull in further sections.

Recently independent proofs of the reversibility property were given in [13] and [32] .
Kawashima [22] generalized this property, stating that also both the distributions of service completion times for every customer and the number of service completions in the time interval /0 , t / are invariant for reordering the servers reversely.

Yamazaki, Sakasegawa and Kawashima [55] proved, that the reversibility property holds also for the case when, at some stages, there are installed homogeneous multiservers, having however constant service times /notice that this is not a queueing line any more/.
Wolisz [52] verified, that this property holds also for two-stage homogeneous systems with arbitrary numbers of exponential servers at each stage.

The occurance of blocking phenomenon causes a decrease of system throughput in comparison with systems without such effect.
In fact Muth [31] pointed out, that the capacity V_L of any queueing line with finite intermediate buffers has two bounds:

- The upper bound V_L^+

$$V_L^+ = \min_{i=1,2,\cdots M} \mu_i \quad , \tag{7}$$

being equal to the service intensity of the slowest server.
The throughput of a queueing line would tend to this value if all queues were allowed to build up without any restrictions.
This is a direct result from Sachs [44] theorems concerning the ergodicity of tandem queueing systems.

- The lower bound V_L^-

$$V_L^- = \frac{1}{E[\max(b_1, b_2, \ldots b_M)} \tag{8}$$

It is easy to see that $V_L^- = V_L^+ = V_L$ in the case of constant service times .

In his paper Muth compared also the values of the difference $V_L^+ - V_L^-$ for various queueing lines.

The capacity of a MQS may be, generally speaking, influenced by the number of stages as well as following parameters describing individual stages:

-service intensity,

-service time distribution,

-buffer size allocation to the stages,

-servers reliability,

-number of servers installed at the stage,

-homogenity of servers installed at one stage.

Notice the common assumption, that a server may break-down only when service is in progress.Thus assuming that the preempted by some breakdown service is resumed after repairing of the proper server we can under some simple additional assumptions, treat the breakdown process together with the service process, describing them jointly with a modified service duration distribution function.Thus if we shall further assume, that some service time distributions are identical that will mean identicity of both the real service and break-down processes.

A MQS is defined by specifying the parameters of all stages which are - generally speaking - entirely different for individual stages. Let us point out that two main types of optimization problems are usually formulated:

a/ the improvement problem

Given an M stage system, increase its capacity through modifying the parameters of some stages.

Usually it is demanded to point out which possible modification / in the context of real process being modelled/ would be most profitable in terms of throughput increase.

b/ the rearrangement problem

Given an M stage system arrange all the facilities so as to maximize its throughput.This problem can be solved through one of the following actions:

- different dividing of processing among stages, thus influencing the mean service times of the stages involved / like for example in the case of two-stand rolling mills/,
- changing the sequence of stages / this is permitted in some processes like for example equipment maintenance,testing or tuning/,
- different allocation of buffers to individual stages with regard to the fixed total buffer size Z.

It was demonstrated / for example using approximate calculations in [19] and simulation in [41]/that MQS capacity decreases generally with the increase of number of stages.Further an attempt is done to present outlines for optimization in both of the above precised contexts with respect to parameters of individual stages.

4. MQS with constant service times .

We shall start our considerations on throughput optimization in MQS with the special type of systems,having constant service times.

Such systems have been considered in papers [9] , [16] , [27] , [31] , [49] .

As the Lavenberg theorem mentioned in the previous section does not directly apply to this case,the open systems have to be considered.

It has been proved, that for any open system of homogeneous type with i-th stage consisting of n_i parallel channels, each having the same constant service time $b_i^* = c_i$, neither the sujourn time of demands nor the stream of demands leaving the system depend on the sequence of stages and the capacity of intermediate queues.This result remains valid also for arbitrary input streams, and in the case of intermediate queues allowed to build up to infinity [49] .

Friedman [16] has introduced the concept of dominance, assuming that stage k dominates stage p if no demand ever waits at stage p if it is preceded by stage k .
He proved also that such property holds if and only if $c_p \leqslant K c_k$, where $K = [n_p/n_k]$ /greatest integer notation/.
For example when one stage dominates all other, then the only waiting in the system occurs before this very stage, which can be easily demonstrated through proper rearrangement.
Using this concept Friedman suggested, while Suzuki and Kawashima [49] developed further, a method for reducing a multi-stage system to an equivalent system with smaller number of stages, sometimes even to one stage succeded by a $\bullet/D/\infty$ system.As the stability condition of the equivalent G/D/n system can be easily established, it is possible in such special cases, to compare using the presented above methodology, the changes of the original system's capacity for different parameters of individual stages, eventually chosing the best one.This method is however of strongly restricted use.

Generally valid results may be found on this basis only for queueing lines.Muth [31] pointed out, that the capacity of such line T is given by (7), (8)

$$T = [max(c_1, c_2, \cdots c_M)]^{-1} \qquad (9)$$

Evidently the maximal capacity is achieved, independently from the buffer size, for the balanced case.

The problem of unreliable servers leads in fact to consideration of systems having variable service times, and as such will be discussed in further sections.

5. Unbalancing of queueing lines .

Let us assume a M stage queueing line with service time distributions $B_i(x)$, $i = 1, 2, \ldots ,M$.

According to (2) we shall assume that

$$\sum_{i=1}^{M} \frac{1}{\mu_i} = M \qquad (10)$$

$$\text{where} \quad \mu_i = \left[\int_0^\infty x \, dB_i(x) \right]^{-1}$$

a/ The case of identical service time distributions.

We shall temporarily constrain ourselves to the case of identical /up to the mean value/ service time distributions.Thus the only parameter which will be changed are the service time intensities μ_i , however with respect to (10) .Also the size of intermediate buffers is assumed to be fixed and equal for all stages: $N_i = N$, $i = 2, 3, \ldots, M$.

For the case of two-stage systems, the reversibility property leads to a conclusion that the balanced case is optimal.

Fig. 2 presents system capacity versus its unbalancing in the case of exponential service time distributions / data taken from [18] /. Notice that losses due to unbalancing increase rapidly with the increase of intermediate buffer size N.Thus proper system balancing becomes more crucial for bigger values of N.

Hillier and Boling [18] investigated also the case of $M = 3, 4$; again for exponential service time distributions.They proved that the balanced case is not optimal any longer, and suggested the existence of so called "bowl phenomenon", asking for higher service intensity being assigned to middle stations.

The optimal solution is "symmetric" / $E(b_1) = E(b_3)$ for M=3 ; $E(b_1) = E(b_4)$, $E(b_2) = E(b_3)$ for M=4 / as it should have been expected due to the reversibility property.Proper unbalancing not necesserily very precise, leads to some gain, while inproper unbalancing leads to significant losses.Some remarkable data are given in Table 1.

It is worth noticing that applying $E(b_2) = 1.15$ for the case M=3,N=0; leads to the capacity equal to 98.7% of the balanced case.

As previously,the system sensitivity to inproper balancing increases with the increase of N.

Patterson [35] analysed three stage systems by means of numerical methods obtaining similar results.He suggested however, that for $M > 3$ the optimal arrangement should be such, where the quick stations separate the slow ones.He suggested also,that the gain obtained from such a procedure should decrease with the decrease of service time variability.

A simulational study of queueing lines with M = 3, 4, 12; $N \geqslant 0$ was presented by El-Rayah [41] who compared the above given strategies with the balanced case.In fact he tested also a third strategy of assigning to the consecutive stations low- medium- high service times, respectively.Such strategy was suggested by Davis [14],for a system with losses /without blocking/, and it was improbable that it will be adequate for the considered case.This strategy is however also suggested sometimes as reasonable.

Experiments with exponential service time fully confirmed the "bowl phenomenon" hypothesis.Furthermore, also for normal *) and lognormal service time distributions the bowl-type arrangement of mean service times was found to be the best one, leading always to improvement over the balanced case.

As for the case M=12 one could suggest various possibilities of defining the bowl - type arrangement, some of them have been tested. The arrangement with two middle stages having the smallest service time, which increased stepwise with equal quantum up to the longest values /symmetrically/ on the ends, was found better than those where "bowl" was formed from groups of 3-4 servers having equal service intensity in every group. This arrangement applied for the exponential service time distributions leads/to the solution given in Table 2 .

The bowl-type unbalancing is always efficient. It was demonstrate that both the possible gain from proper unbalancing, and the imbalance itself, increase with : greater variability in operation times, smaller interstage queueing capacity and larger number of stages in the line. The gain in capacity, possible to achieve due to unbalancing is never high. The possible gain for M=12 /which perhaps could be a little bit improved chosing non-identical quanta while unbalancing the system/ is only marginally higher then that for M=4.

The unbalancing method suggested by Patterson is very unreliable, and leads frequently to system capacity lower then the balanced case.

The low-medium-high arrangement was definitely found to fail generally, and was significantly worse /as a rule/ than the balanced case.

The study of El-Rayah, based on solid statistical methods, shows however how one should be cautious in assessing the results of simulation. Mean values of capacity for dual systems were usually different but the difference was assessed to be statistically insignificant - a correct result in view of the reversibility property. Curiously enough for the three stage system /p.66 of [41] / the author state "It was also verified that a low-medium-high arrangement is superior to a high-medium-low arrangement in terms of expected output rate" which is obviously wrong!.

*) Naturally here and furtheron the trunctated normal distribution is considered, (p 15).

b/ The case of different service time distributions.

Rao [39] presented a methodology for establishing the capacity of two-stage lines with exponential service time distribution at one of the stages, and general service time distribution at the other stage, calling for solution of a system of linear equations which order was dependent on the buffer size N.

Calculations for Erlang and normal distribution with variability coefficient $C < 1$ lead to the conclusion that system capacity is optimized if the server having greater variability is assigned slightly higher speed. Both the gain and optimal imbalance increase with increase of the difference in variability coefficients and decrease significantly for larger values of N.

Similar systems have been considered by Wolisz [52] where closed form expressions for system capacity have been given. The earlier results for $C < 1$ have been confirmed, but surprisingly quite different observations appeared for $C > 1$, investigated with second - order hyperexponential distribution. The optimal capacity was in this case obtained for slightly quicker exponential server, which however this time was the one having smaller variability coefficient. Also the increase of N led initially to the increase of gain /and unbalancing/, and only further increase of N reduced this effects.
Sample results are plotted in Fig 3.

Typical optimal unbalancing in the two-stage line lies in the range $E(b_1) = 0.92 - 0.96$ leading to a gain of 0.2 - 0.3 % over the balanced case.

Rao [40] investigated also analytically the case of M=3 without intermediate buffers, assuming all possible combinations of exponential and deterministic service time distributions.

Table 3 contains the results of optimal unbalancing /notice the existence of dual systems/. Rao introduced the term "variability imbalance" calling for assigning shorter service times to the more variable stations.The "variability imbalance" effect may either coincide /eg pattern f/ or contradict /cf pattern c/ with the "bowl phenomenon".

It was suggested that the strength of the variability imbalance effect depends on the difference in variabilities.Rao verified that if a server with uniform service time distribution and variability coefficient equal to 0.5 is located between two exponential servers then the balanced case becomes optimal.

Concluding we can establish the existence of both the "bowl phenomenon" and the "variability imbalance",effect of the later being clear for $\varsigma < 1$,while the case of $\varsigma > 1$ needs further research. Joint effect of the two phenomena given above can be significant /cf the 6.79% gain in patterns d,e,f of Table 3/.

6. <u>The effect of service time distribution on system capacity</u> .

The irregularity of service time is caused by two main reasons:

- the inavoidable stochastic differences among demands as well as stochastic disturbances in the service process.Those lead usually to small changes of service time - lying in the range of variabilit coefficients less than unity / or more frequently less than 0.5, as for example a value of $\varsigma = 0.27$ was found typical for the production lines by Slack [47] .Values of ς approaching one are reported in some data transmission applications/.
- break-down of the server.Such situations occure rarely, but it takes usually a mean repair period several times longer than mean service time to resume server's operation.Thus the resulting joint service time has in this case a high variability.

The effects of those two reasons for service irregularity are usually studied seperately : either perfectly reliable servers are assumed or the service times are assumed to be constant, while in random periods break-downs of random duration are assumed.

In queueing systems for the sake of simplicity, there is a strong trend to characterize the random variables involved, only with two first stochastic moments.Let us first assess what error does such attitude introduce in the case of MQS.

a/ The effect of ignoring service time distribution higher moments.

Fig 4, based on data from [38] compares system capacity for M=2,N=0 versus C for different /Erlang and normal/ service time distributions at both stages, suggesting that the differences are of quantitative type only, and increase with the increase of C . Rao [39] demonstrated, that this effect becomes even more visible when quite different types of distributions are compared, like Erlang and uniform distributions with identical values of C .He provided also examples that if, in a two-stage system one of the stages has some fixed distribution, then with the increase of its variability coefficient the influence of higher moments at the other stage will be greater.This influence becomes stronger in the case of small intermediate buffer /Table 4/.

Anderson and Moodie [1] used in a simulation study of multistage balanced lines aiming at optimization of buffer size from the point of view of some complex criterion /cf section 9/ both the exponential and normal service time distributions, finding the results qualitatively identical and quantitatively similar, however different.

Such conclusions were also presented by El Rayah [41] who used both normal and lognormal distributions in his simulation studies of unbalanced queueing lines with M=3,4.

Thus we conclude, that for $C < 1$ no example of qualitative difference in optimization rules due to the higher moments are known, while conditions when the quantitative differences can be expected to be small where listed above. In practical cases, for simulational studies where the choice of distribution functions is unrestricted, some afford is made to preserve the shape of this function, using for example positively skewed Weibull distributions, eg [10] , [47] .

b/ The influence of service time variability on system capacity.

From Fig. 2,3,4 and Tables 3,4 it is evident, that system capaci decreases with the increase of service time variability coefficient at any stage. This loss becomes however significantly smaller when larger buffers are applied. Similar conclusion was achieved by Barten [24] for a simulation study of six-stage systems with normal service time distributions, who stressed specially the beneficial role of buffers i canceling the bad influence of service time variability. Data supportin this property can be found also in a simulation study with Weibull type distributions [10] .

On the other hand it has been observed that a similar effect is visible in queueing lines with constant service times and unreliable servers. Studies of such systems where reported [7],[33] ,[34] for paced queueing lines. Buzacott [9] demonstrated however their direct applicability for the unpaced case as well. He investigated analyticall a balanced two-stage system with $N \geqslant 0$ and exponential service times, in which stochastic breakdowns of random duration occured.

It was pointed out that the decrease of system capacity in comparison with the fully reliable - constant service time case, can be very precisely approximated by a sum of losses due to either service time variability itself or nonperfect reliability.

Such superposition was expected to hold also for larger systems.

/ Arrangement of balanced queueing lines due to service time variability.

Some data on a balaneed three-stage queueing line with different rdering of servers having unequal variability coefficients are given s a result of a simulation study by Smith and Brumbaugh [46] . hey suggested that allocation of a station with highest variability n the middle of the line led to the worst results,as everage calculated over different buffer allocation patterns.The data presented n Table 1 of [46]/in terms of means only/ violate however significantly the reversibility property, thus it is not possible to draw sing them any more detailed conclusions.

Systems with M=4,10 and different values of N have been simulated y Carnall and Wild [10] .Service times where assumed to be either onstant or variable, described by the positively skewed Weibull type istribution.In every experiment it was essumed that variable stations re identical.

It was found that for M=4, and two variable stations significantly igher capacity has a system with variable stations located at the nds of the line, in comparison with the case of their location at the iddle. The gain from such strategy increases with the increase of ariability coefficient and decreases with the increase of buffer sizeN. or example in the case $C = 0.5$, M=1 the gain approaches 4% .

Experiments with M=10 lead also to a conclusion that locating ariable stations at the ends is justified, resulting in a 1.33% gain ver a random sequencing of stages, and 3% gain over allocation of ariable stages in the middle of the line.

Thus the authors suggested the existence of something like the bowl phenomenon" concerning the service time variability for balanced ines.

An extensive simulation study with normally distributed service times and no intermediate storage was reported by El-Rayah [42] . For M=3 the bad effect of allocating highest variability to the middle station was confirmed.Also for M=4 the above given suggestions were verified to be true.Further the author demonstrated that this queueing line with equal mean service times and "unbalanced" variability coefficients /the sum of them over all servers being constant/ may yield higher capacity then a line with identical servers. It was demonstrated for this case, and also for M=12 that something like the "bowl phenomenon" exists also for the service time variability In the investigated range of $C < 0.3$ it was found that increasing of some servers variability coefficient, and decreasing its mean service time yield highly similar results.For example if 4 out of 12 servers had $C = 0.15$, while other had $C = 0.3$ then locating the servwrs with smaller variability in the middle of the line, instead of locating them / in one group/ either at the beginning or at the end of the system /both of these cases being equivalent/ led to an improvement in capacity of over 2.5%.

7. The effect of intermediate buffers on system capacity .

In previous sections it was several times mentioned, that the intermediate buffers may increase or decrease the above discussed effects.Now we shall discuss directly the influence of intermediate buffers on system capacity.

In section 3 it was mentioned that intermediate buffers eliminate / to some extent/ the blocking and idleness of individual stages. Thus it is clear that both the preceding and consecutive buffer size influence the operation of any stage.

urthermore as it yields from (7) that for the unlimited intermediate uffers always the balanced system yields the maximal throughput, we onclude /cf [7] ,[31] / that buffers cannot reduce that portion of roduction line inefficiency which is caused by unequal mean service imes at different stations.

Buzacott [8] points out, that if due to long term imbalance etween stages a buffer is permanently full / or permanently empty/ t is serving no usefull purpose.Thus the magnitude of queue length ariations may serve as a measure of the buffer effectivness. hese remarks are consistent with results cited in section 5, where or large enough buffer sizes the balanced case was demonstrated to e optimal.

Certainly as it is visible from the previous section, buffers may ignificantly decrease the bad results of service time irregularities.

Using the upper bound V_L^+ given by (7) , one can suggest, that he ratio V_L / V_L^+ is some measure of buffer efficiency.Let us notice hat in all the figures and tables presented so far $V_L^+ = 1$.

Fig. 2 gives a good example of buffer size N influence on system apacity.This influence can be presented in a simple, analytical form. unt [20] found that for a two-stage balanced system with exponential ervice times and $E(b_i) = 1$; $i=1,2$ system capacity can be expressed y a simple formula

$$V = \frac{N+2}{N+3} . \qquad (11)$$

hus adding an additional waiting place to a buffer of size N leads o a relative gain of E,

$$E = \frac{1}{N^2+6N+8} . \qquad (12)$$

alues of both V and E for different N are printed in Table 5.

t is evident that increasing the buffer size by one leads always to some gain being however significant for small N and only marginal for large N.

Thus Hunt concluded that using buffers of size larger than 5 doesn't, generally, pay.

Similar conclusions can be drown for unbalanced two-stage queueing lines with nonexponential servers from Fig 3, and for three-stage exponential lines from Table 1.

Barten [4] simulated systems with normal service time distributions, M=4,6,10 and different values of variability coefficient, demonstrating that the above given remarks remain generally true. This was also confirmed by Slack and Wild [48] for M=5,10,15.

Evidently, however, inorder to achieve some predetermined system capacity, the buffer sizes applied should be larger, if service time variability increases /cf Table 4/.
The situation changes significantly when servers break-downs are included into consideration.
In this case, as pointed out by Buzacott [7], the minimal size of buffer should be equal to the mean number of services completed during the mean repair time, while 2-3 times larger buffers seem to be thereasonable choice.The typical buffer size would be rather 30-50 this time.

The problem of allocating the proper joint buffer size Z to a queueing line and dividing it inbetween different stages obtained a lot of attention, eg. [15], [17], [23], [25], [26], [29] . Unfortunately enough the majority of research was devoted to paced queueing lines and thus the obtained results are not of direct use in our case.Let us present an example of the differences.

For paced lines it is stressed / [15], [29] / that the unreliable stages /in other words :stages with high variability/ located at the end of the line ask for more significant increase of buffer size in order to achieve proper system capacity, than identical stages positioned in front of the line.This is obviously in disagreement with the reversibility property.Also the suggestion / [15] / that the division of total buffer size Z among different locations in the line doesn't depend on the value of Z is not applicable for the unpaced case.

An extensive study of buffer allocation problems has been presented by El-Rayah [42]. The case of balanced queueing lines with M=3,4 has been simulated, assuming identical variability coefficients of all servers. The objective of experiment was to verify, should the intermediate queues have identical maximal sizes, or should they be unequal, preserving during the experiment the condition

$$\sum_{i=2}^{M} N_i = Z = const.$$

It was verified, that for $Z \leq (M-1)4$ indeed equal buffers yield the maximal system capacity, as what could have been anticipated - every attempt to decrease any of the buffers significantly handicappes the stages involved.

For larger values of Z it occured that the equal assignment leads to almost optimal results; sometimes only assigning larger buffers to middle stages yields some - almost negligible - gain.

Thus it occurs that increasing the buffer size at some stations leads to similar, but considerably smaller, effect as increasing the servers speed.

This effect being small enough does hardly lead to "unbalancing" queueing lines with identical servers in the sense of buffer capacities, it may however be quite valuable when different servers are involved.

Smith and Brumbaugh [46] concluded, that service time variability should be considered when allocating buffer sizes. They stated that relatively greater buffers should be allocated around more variable stages, while possible benefits from proper allocation are smaller than losses incurred if the arrangement is improper. Departures from equal buffer allocation were found to have grater impact when the total buffer size Z was small.

8. The effect of introducing parallel servers .

In sections 5-7 only the case of queueing lines was discussed. Let us now consider the effect of introducing multiservers at some stages.

Wild and Slack [51] in a simulation study compared the operation of two queueing lines with the case of a MQS having two servers at each stage and the same buffer size per server.It was found that the second system is always more effective, the gain being high in the case of large number of stages, low buffer size,and high service time variability, and comparatively lower otherwise.

A systematic analytical study of homogeneous two-stage queueing systems with exponential multiservers at both stages was published by Wolisz [53] .

a/ Two-stage systems : the balanced case.

For the balanced case /as defined by equality (3) / the influence of buffer size N and number of servers $n_1 = n_2 = K$ on system capacity was investigated.As presented in Fig.5 the influence of intermediate buffer size is similar as in the queueing lines.It is worth stressing, that system capacity increases with the increase of the parallel servers number K, as the additional servers act also as additional buffers.Thus in the case when the buffer capacity is strongly limited, applying of many slow servers at every stage pays better then using a few quick ones.

Also another experiment was reported there.The total system accumulation L was assumed to be constant, it could however be differently devided between servers and buffer with respect to an equality

$$2K + N = L$$

Results for $L \leqslant 10$ are plotted in Fig.6 showing that it is evidently better to have as large buffer as possible.Furthermore a system with larger accumulation can have a lower capacity than another system having smaller accumulation, but devided so as to favor buffers size on expence of the number of servers / cases K=4,N=2 and K=2,N=4 compared, may serve as example/.

Thus the following rule should be applied:
If possible a queueing line with large buffers should be used.
If there is no possibility of introducing buffers / or their size is severly limited/ then the use of slow multiservers instead of quick single-servers results in capacity increase.

Fig 7 demonstrates that systems having the equal number of servers at both stages /buffer size being constant/ are always most efficient.

A comparison of a MQS and several queueing lines as presented in Fig 8 was done resulting in an experiment similar to that reported in [51] .The results are presented in Table 6.
The conclusions of[51] have been fully supported.Thus the system from Fig 8b was always better .

b/ Two-stage systems - the unbalanced case

Let us denote for the sake of simplicity

$$\mu_I = n_1 \cdot \mu_1^* , \quad \mu_{II} = n_2 \cdot \mu_2^*$$

According to (2) following equation is to be respected

$$\frac{1}{\mu_I} + \frac{1}{\mu_{II}} = 2$$

System capacity for $n_1 = n_2 = K$, $N = 0$ is plotted in fig 9, showing that in this case the balanced system is always optimal.Losses due to unbalancing increase with the increase of K.

If however the number of servers at both stages is unequal, then proper unbalancing i.e. assigning higher service intensity to the stage having larger number of servers leads to increase of system capacity.

Examples advocating this statement are plotted in Fig 10, where a constant difference in numbers of servers $n_1 - n_2 = 2$ was assumed.

Both the gain over balanced case and optimal unbalancing decrease with increase of either buffer size N or system accumulation L, and increase with increase of the difference $n_1 - n_2$.

However it was demonstrated that although for $n_1 \neq n_2$ an optimally unbalanced system has higher capacity than a balanced one, it is worse than a balanced case where the identical total number of servers $n_1 + n_2$ would be equally devided between the stages.The loss is higher for small $n_1 + n_2$ and small N.

Thus to optimize the capacity of such systems one should try to apply equal numbers of servers at both stages and balance the system, only if this is impossible, the loss resulting from the unequal numbers of servers may be minimized by proper unbalancing.

9. Remarks on other optimization criteria .

Throughout this paper we were concerned with a unique optimization goal : system capacity maximization.In this section we shall mention briefly other characteristics which are also frequently considered as important design factors.

The most commonly used characteristics are:

- The expected number of demands in the system
- The idle time of individual stages
- The mean in- system time /sujourn time/ of demands.

It should be mentioned that those characteristics are of significant interest both for open and saturated systems, as well as for systems with infinite intermediate queues.

It has been proved / for example [42] ,[46] / that optimal system parameters chosen for different of this characteristics do not coincide.Also the sensitivity of this factors on parameter changes is quite different.

For example for e queueing line with M=3,Z=8 changing the buffer allocation pattern from N_2 =3, N_3 =5 to N_2 =5, N_3 =3 increases the expected number of ùnits in system from 6.3 up to 7.6 not influencing the system capacity at all - systems being dual [42] .

The buffer allocation from small to large along the line was advocated, as decreasing significantly the expected number of units in system,with only marginal loss of capacity for larger M as well.

Optimal sequencing of open two-stage queueing lines minimizing the delay was studied analytically by Tembe and Wolff [50] for infinite, and by Kawashima [22] for finite intermediate queues.

Simulation studies of queueing lines with infinite intermediate queues together with optimization outlines were reported for M=2 in [30] and [45] while lines with M=4 and M=20 were studied in[21][30] respectively.

Complex optimization criteria are often used for determining the proper buffer size [5], [1], [23] .

It is suggested that buffer size should be chosen with regard to following factors:

- costs of servers idle time
- costs of decreasing systems throughput due to limited storage
- costs of holding the inventory of demands in system
- cost of storage space

Closer discussion of these problems exceeds however the scope of this paper.

10. Conclusions .

In view of the statesments presented above, some general conclusions can be made.

Let us introduce informally the concept of individual stage efficiency described as following: if the efficiency of any stage increases, while other stages remain unchanged, then the capacity of the whole system will be increased. We shall also assume that the efficiency of identical stages is equal.

The reviewed in this paper results demonstrated, that the efficiency of individual stages may be increased by:

- decreasing service time / i.e. increasing service intensity/,
- reducing service variability ,
- improving servers reliability,
- increasing the size of buffer belonging to this stage ,
- replacing a single server with a multiserver with identical service intensity.

Generally queueing line achieve the maximal throughput, when the stages efficiency is unequal, most efficient stages being located in the middle of the line, and less efficient being gradually moved in the direction of its ends. This can be, essentially, obtained by changing any of the above listed parameters, however with quite different sensitivity. Some illustrative data were given in previous chapters.

It should be strongly stressed, that the value of results review here lies by far more in their qualitative than quantitative part. In practice neither obtaining strictly equal nor optimally unequal stage parameters is possible with high precision.

The most important thing is to know, in which direction should the unevoidable discrepences from the ideal case be permitted, in order t cause serious losses.

Notice that in all the optimization outlines, as a rule the possible gain from observing the optimization pattern was significantly smaller than the loss caused by a wrong decision , although sometimes the gain itself was also worth approaching .
In this sense the exact knowledge of service time distribution is not essential, as long as their variability coefficient does not exceed unity.All the results remain qualitatively similar in this case, and quantitative discrepences are,usually, small.

The case of $C > 1$ received very few attention so far. This occurs if the breakdown process is treated jointly with the service Also in the case if at some stages one of several non-identical servers is used alternatively for a given demand / because of its special, individual features/ such situation occurs, leading for example to hyperexponential service time distribution.
Such situations are not uncommon while modelling production systems. As some irregularities have been noticed /cf. section 5/ for $C > 1$ this area needs further research.

Similarly further research is needed in order to verify to what extent the results of introducing parallel servers discussed in section 8 may be generalized for the system with larger number of stages.

It is essential while using optimization rules suggested in some papers to make sure, that the models are quite identical in the referred report and in the application considered.
Misunderstandings caused by some rules suggested for paced systems in application to the unpaced case where mentioned earlier.
Similarly the Davis [14] conjecture about unbalancing pattern is frequently cited without noticing that it was orginally formulated for a system with losses /eg [10] / !

Seemingly small differences in the investigated models lead sometimes to quite different optimization rules.

As it was illustrated by figures and tables the gain in capacity obtained through system optimization is sometimes small.
Thus there is an essential difficulty in optimization studies for systems where no exact analytical results are available / the error introduced by approximate methods may deform the conclusions obtained/

Using simulation,a great afford must be done to verify the statistical significance of the obtained results, erroneous statesments being by far not uncommon.

In this paper figures and tables were constructed mainly for the models investigated by use of analytical tools, to avoid quoting some mean values, being meaningless without citing also the backing them statistical reasoning.

Finally it has to be stressed that factors other than system capacity generally react differently to the optimization rules presented here.However the information about effects of individual stages parameter changes on system capacity remains important also if other optimization goals are chosen, making it possible to decide on a reasonable policy for the occuring tredeoffs.

References

[1] Anderson,D.R.,Moodie,C.L., "Optimal Buffer Storage Capacity in Production Line Systems ", International Journal of Production Research 7 /1969/ pp 233-240

[2] Avi-Itzhak,B.,Yadin,M., "A Sequence of Two-Servers with No Intermediate Queue" Management Science 11 /1965/ pp 553-563

[3] Avi-Itzhak,B., "A Sequence of Service Stations with Arbitrary Input and Regular Service Times " Management Science 11 /1965/ pp 565-571

[4] Barten,K., "A queueing Simulator for Determining Optimum Inventory Levels in a Sequential Process" The Journal of Industrial Engineering 13 /1962/ pp 245-252

[5] Basu R.N., "The Interstage Buffer Storage Capacity of Non-Powered Assembly Lines, a Simple Mathematical Approach", International Journal of Production Research, 15 /1977/ pp 365-382

[6] Buxey,G.M.,Slack,N.D.,Wild,R. "Production Flow Line System Design-A Review " AIIE Transactions 5 /1973/ pp 37-48

[7] Buzacott,J.A., "Automatic Transfer Lines with Buffer Stocks " International Journal of Production Research 5 /1967/ pp 183-200

[8] Buzacott,J.A., "The Role of Inventory Banks in Flow-Line Production Systems" International Journal of Production Research 9 /1971/ pp 425-436

[9] Buzacott,J.A., "The Effect of Station Breakdowns and Random Processing Times on the Capacity of Flow-Lines with In-Process Storage" AIIE Transactions 4 /1972/ pp 308-312

[10] Carnall,C.A.,Wild,R., "The Location of Variable Stations and the Performance of Production Flow Lines " International Journal of Production Research 14 /1976/ pp 703-710

[11] Caseau,P.,Pujolle,G., "Throughput Capacity of a Sequence of Queues with Blocking Due to Finite Waiting Room" Typescript/1978/ to appear in IEEE Transactions on Software Engineering

[12] Chang,W., "Sequential Server Queues for Computer Communication Systems Analysis" IBM Journal of Research and Developement /1975/ pp 476-485

[13] Dattatreye,E.S., "Tandem Queueing Systems with Blocking" Ph.D. Dissertation, IEOR Department, University of California, Berkeley, January 31, 1978

[14] Davis, L.E., "Pacing Effects on Manual Assembly Lines", International Journal of Production Research 4 /1966/ pp 171-184

[15] Freeman, M.C., "The Effect of Breakdowns and Interstage Storage on Production Line Capacity" The Journal of Industrial Engineering 15 /1964/ pp 194-200

[16] Friedman, H.D., "Reduction Methods for Tandem Queueing Systems" Operations Research 13 /1965/ pp 121-131

[17] Hatcher,J.M., "The Effect of Internal Storage on the Production Rate of a Series of Stages Having Exponential Service Times" AIIE Transactions 1 /1969/ pp 150-156

[18] Hillier,F.S.,Boling,R.W., "The Effect of Some Design Factors on the Efficiency of Production Lines with Variable Operation Times" The Journal of Industrial Engineering 17 /1966/ pp 651-658

[19] Hillier,F.S.,Boling,R.W., "Finite Queues in Series with Exponential or Erlang Service Times - A Numerical Approach" Operations Research 15 /1967/ pp 286-303

[20] Hunt,G.C., "Sequential Arrays of Waiting Lines" Operations Research 4 /1956/ pp 674-683

[21] Kala,R.,Hitchings,G.G., "The Effects of Performance Time Variances on a Balanced Four-Station Manual Assembly Line" International Journal of Production Research 11 /1973/ pp 341-353

[22] Kawashima,T., "Reverse Ordering of Services in Tandem Queues" Memoirs of The Defense Academy Japan 15 /1975/ pp 151-159

[23] Knott,A.D., "The Inefficiency of a Series of Work Stations - A Simple Formula" International Journal of Production Research 8 /1978/ pp 109-119

[24] Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks" Journal of ACM 21 /1974/ pp 316-328

[25] Koenigsberg,E., "Production Lines and Internal Storage", Management Science 5 /1956/ pp 410-433

[26] Kraemer,S.A.,Love,R.F., "A Model for Optimizing the Buffer Inventory Storage Size in a Sequential Production System" AIIE Transactions 11 /1970/ pp 64-

[27] Labetoulle,J.,Pujolle,G., "A Study of Queueing Networks with Deterministic Service and Application to Computer Networks" Acta Informatica 7 /1976/ pp 183-195

[28] Lavenberg,S.S "Stability and Maximum Departure Rate of a Certain Open Queueing Networks Having Finite Capacity Constraints", RAIRO- Informatique 12/1978/ pp 353-370

[29] Masso,J.,Smith,M.L., "Interstage Storages for Three Stage Lines Subject to Stochastic Failures" AIIE Transactions 6 /1974/ pp 354-358

[30] McGee,G.R.,Webster,D.B., "An Investigation of a Two-Stage Production Line with Normally Distributed Interarrival and Service Time Distributions" International Journal of Production Research 14 /1976/ pp 251-261

[31] Muth,E.J., "The Production Rate of a Series of Work Stations with Variable Service Times" International Journal of Production Research 11 /1973/ pp 155-169

[32] Muth,E.J., "The Reversibility Property of Production Lines" Management Science 25 /1979/ pp 152-158

[33] Okamura,K.,Yamashina,H., "Analysis of the Effect of Buffer Storag Capacity in Traffic Line Systems" AIIE Transactions 9 /1977/ pp 127-135

[34] Okamura,K.,Yamashina,H., "Justification for Installing Buffer Stocks in Balanced and Unbalanced Flow-Line Production Systems" Typescript /1978/, to appear in AIIE Transactions

[35] Patterson,R.L., "Markov Prosses Occuring in the Theory of Traffi Flow Through an N-Stage Stochastic Service System" The Journal of Industrial Engineering 15 /1964/ pp 188-193

[36] Payne,S.,Slack,N.,Wild,R., "A Note on the Operating Characteristic of Balanced and Unbalanced Production Flow Lines" International Journal of Production Research 10 /1972/ pp 93-98

[37] Pujolle,G., "The Influence of Protocols on the Stability Conditions in Packet-Switching Networks" Typescript /1978/, to appear in IEEE Transactions on Communication

[38] Rao,N.P., "On The Mean Production Rate of a Two-Stage Production System of The Tandem Type" International Journal of Production Research 13 /1975/ pp 207-217

[39] Rao,N.P., "Two- Stage Production Systems with Intermediate Storage AIIE Transactions 7 /1975/ pp 414-421

[40] Rao N.P., "A Generalization of the "Bowl Phenomenon" in Series Production Systems" International Journal of Production Research 14 /1976/ pp 437-443

[41] El-Rayah,T.E., "The Efficiency of Balanced and Unbalanced Production Lines" International Journal of Production Research 17 /1979/ pp 61-75

[42] El-Rayah,T.E., "The Effect of Inequality of Interstage Buffer Capacities and Operation Time Variability on the Efficiency of Production Line Systems" International Journal of Production Research 17 /1979/ pp 77-89

[43] Reeve,N.R.,Thomas,W.H., "Balancing Stochastic Assembly Lines" AIIE Transactions 5 /1973/ pp 223-229

[44] Sachs,J,, "Ergodicity of Queues in Series" Annales of Mathematical Statistics 31 /1960/ pp 579-588

[45] Shimshak, D.G., "Optimal Sequencing of Servers in the Two-Station Series Queueing System" Journal of the Operational Research Society 29 /1978/ pp 779-788

[46] Smith,L.D.,Brumbaugh,P., "Allocating Inter-Station Inventory Capacity in Unpaced Production Lines with Heteroscedastic Processing Times" International Journal of Production Research 15 /1977/ pp 163-172

[47] Slack,N.D.C., "The weibull Distribution and Its Use in Describing Service Time Distributions" Prod.Mgmt.Res.Group Monograph 72/1, University of Bradford /1972/

[48] Slack,N.D.C., Wild,R., "Production Flow Line and "Collective" Working: A Comparison" International Journal of Production Research 13 /1975/ pp 411-418

[49] Suzuki,T.,Kawashima,T., "Reduction Methods for Tandem Queueing Systems" Journal of Operational Research Society of Japan 17 /1974/ pp 133-144

[50] Tembe,S.V.,Wolff,R.W., "The Optimal Order of Service in Tandem Queues", Operations Research 22 /1974/ pp 824-837

[51] Wild,R.,Slack,N.D., "The Operating Characteristics of "Single" and "Double" Non-mechanical Flow Systems" International Journal of Production Research 10 /1972/ pp 139-146

[52] Wolisz,A., "More About Two-Stage Production Systems with Finite Intermediate Storage" Report ZSAK-BO /79, Polish Academy of Sciences, Dept. of Complex Automation Systems /1979/, to appear in the International Journal of Production Research

[53] Wolisz,A., "Throughput Optimization of Two-Stage Queueing Systems with Finite Intermediate Buffer" Podstawy Sterowania 10/1980/

[54] Yamazaki, G., Sakasegawa, H., "Properties of Duality in Tandem Queueing System " Ann. Inst. Stat. Math. 27 /1975/ pp 201-212

[55] Yamazaki,G.,Sakasagawa,H.,Kawashima,T., Production Rate Estimated with Flow-Shop Reversibility" Bulletin of the JSME 21 /1978/ pp 161-171

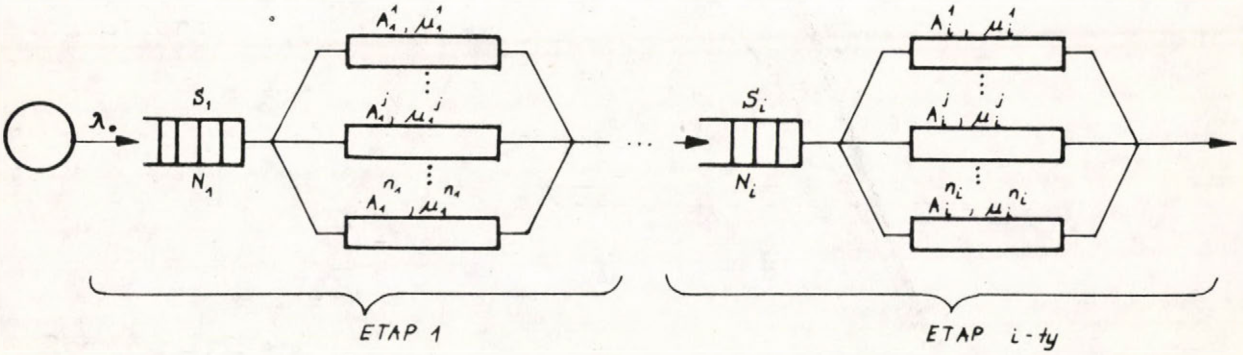

Fig. 1. *A multistage queueing system*

λ_0 - *the intensity of demands arrivals*

n_i - *the number of servers installed at i-th stage*

A_i - *j-th server belonging to i-th stage*

S_i - *buffer for demands waiting to be served at i-th stage*

N_i - *the size of buffer*

μ_i - *service intensity of server* A_i^j

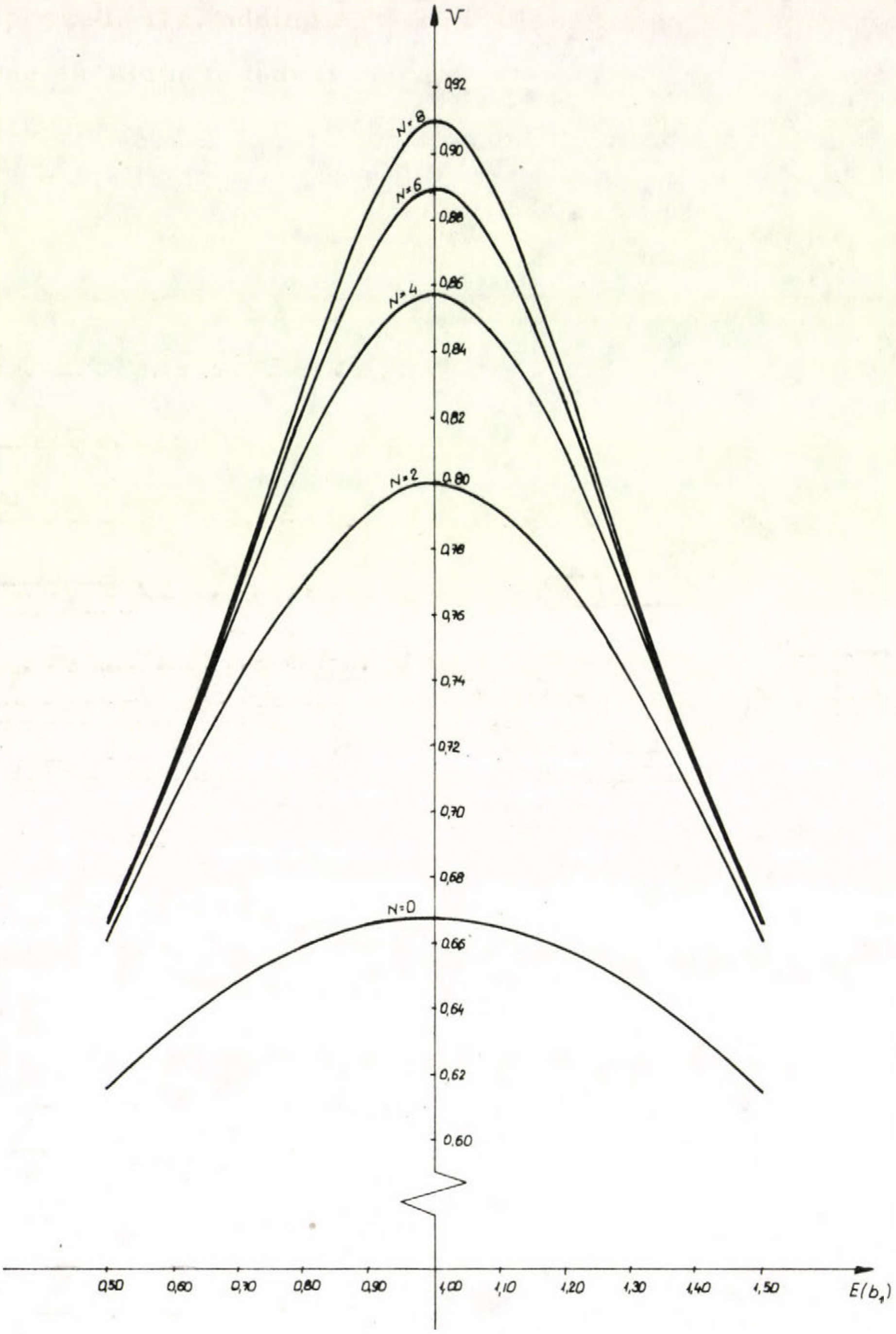

Fig. 2. Capccity of a two-stage unbalanced queueing line with exponential servers. Data taken from [18].

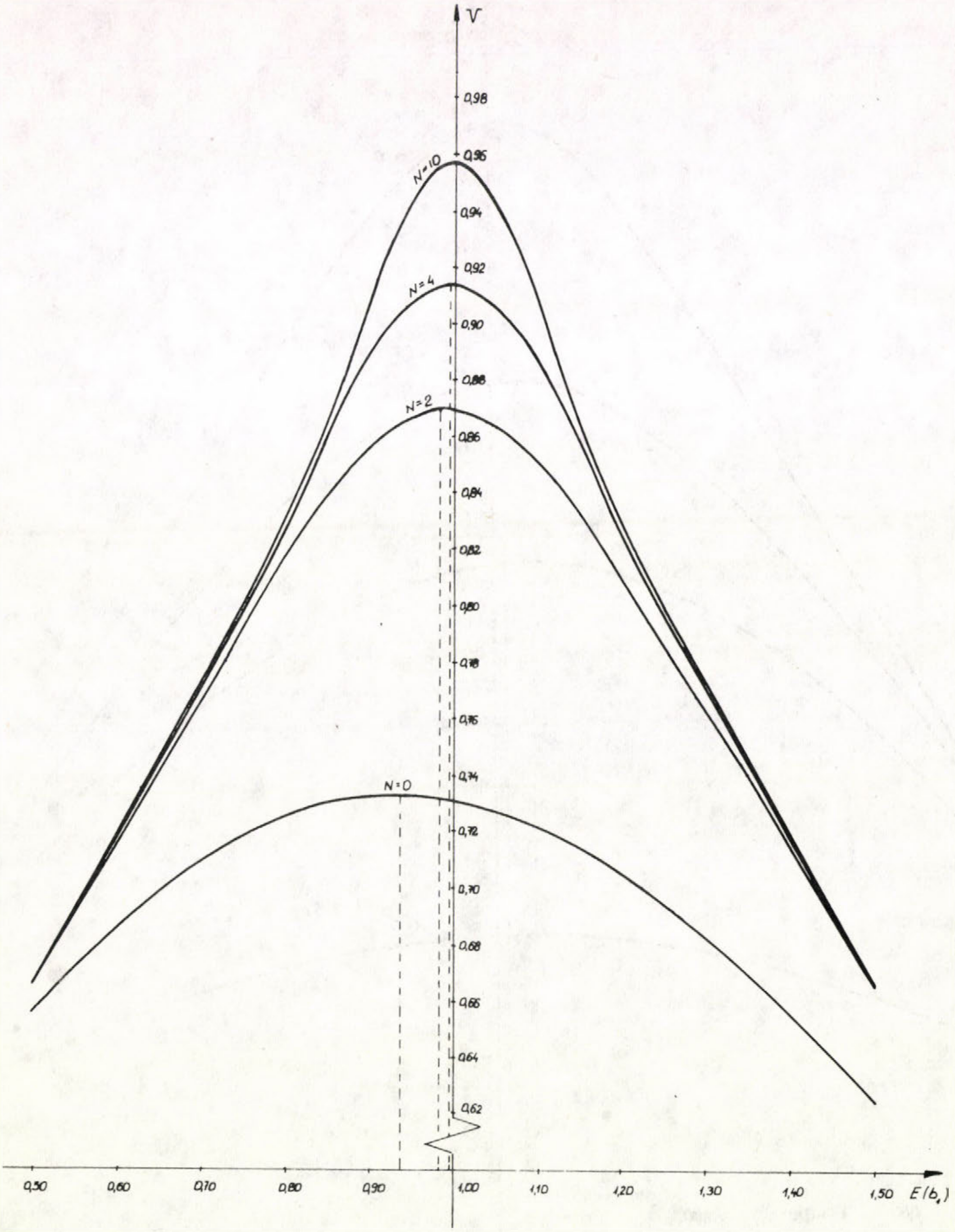

Fig. 3. *Capacity of a two-stage, unbalanced queueing line. Exponential service time was assumed at the first stage, while either regular (case a) or hyperexponential (case b) distribution with $C = \sqrt{3}$ is applied at the second stage. Data taken from* [52].

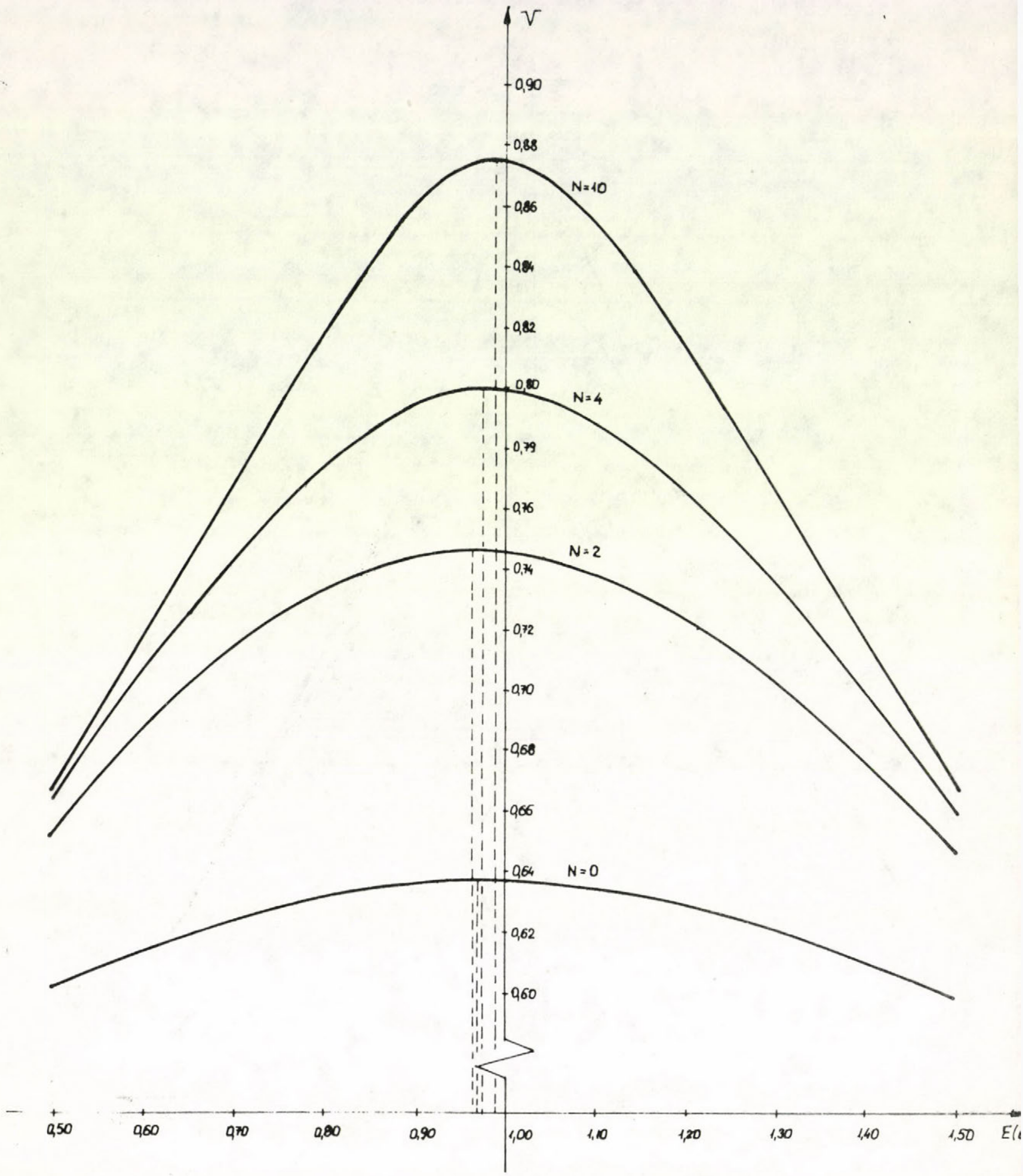
V
0,90
0,88
N=10
0,86
0,84
0,82
0,80
N=4
0,79
0,76
N=2
0,74
0,72
0,70
0,68
0,66
0,64
N=0
0,62
0,60
0,50
0,60
0,70
0,80
0,90
1,00
1,10
1,20
1,30
1,40
1,50
E(

Fig. 3.b

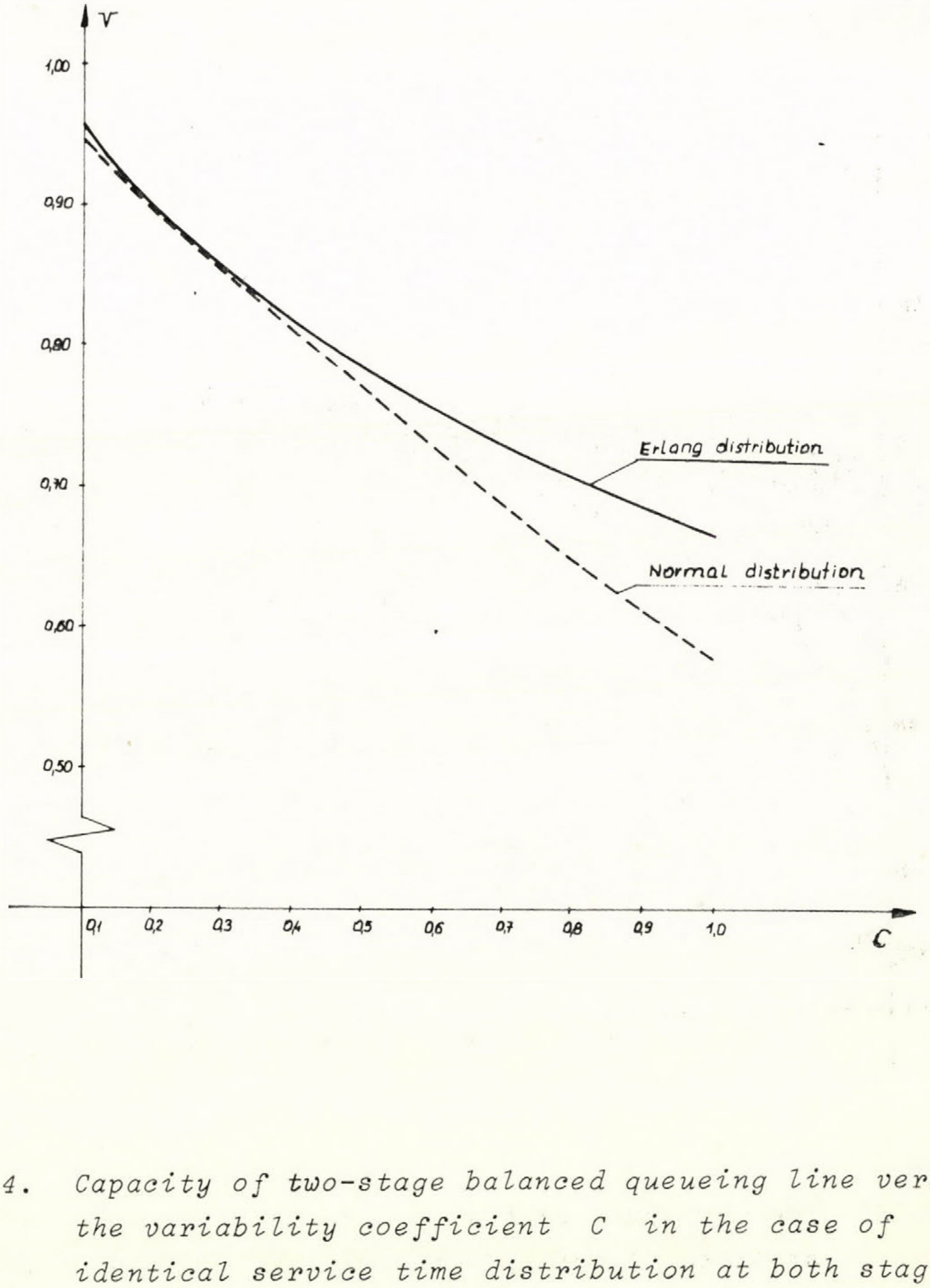

Fig. 4. Capacity of two-stage balanced queueing line versus the variability coefficient C in the case of identical service time distribution at both stages. N = 0. Data taken from [38].

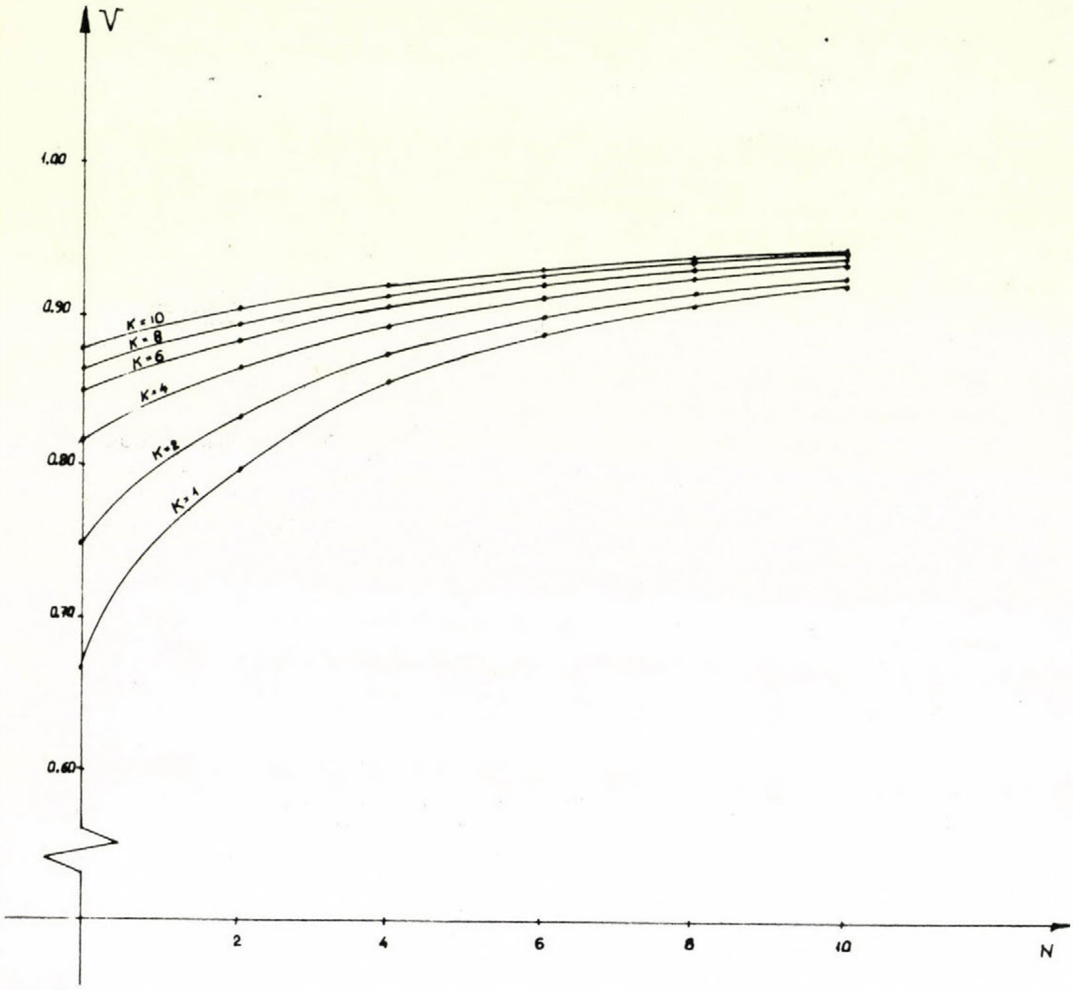

Fig. 5. The influence of buffer size N and number of stations K on system capacity. Data taken from [53].

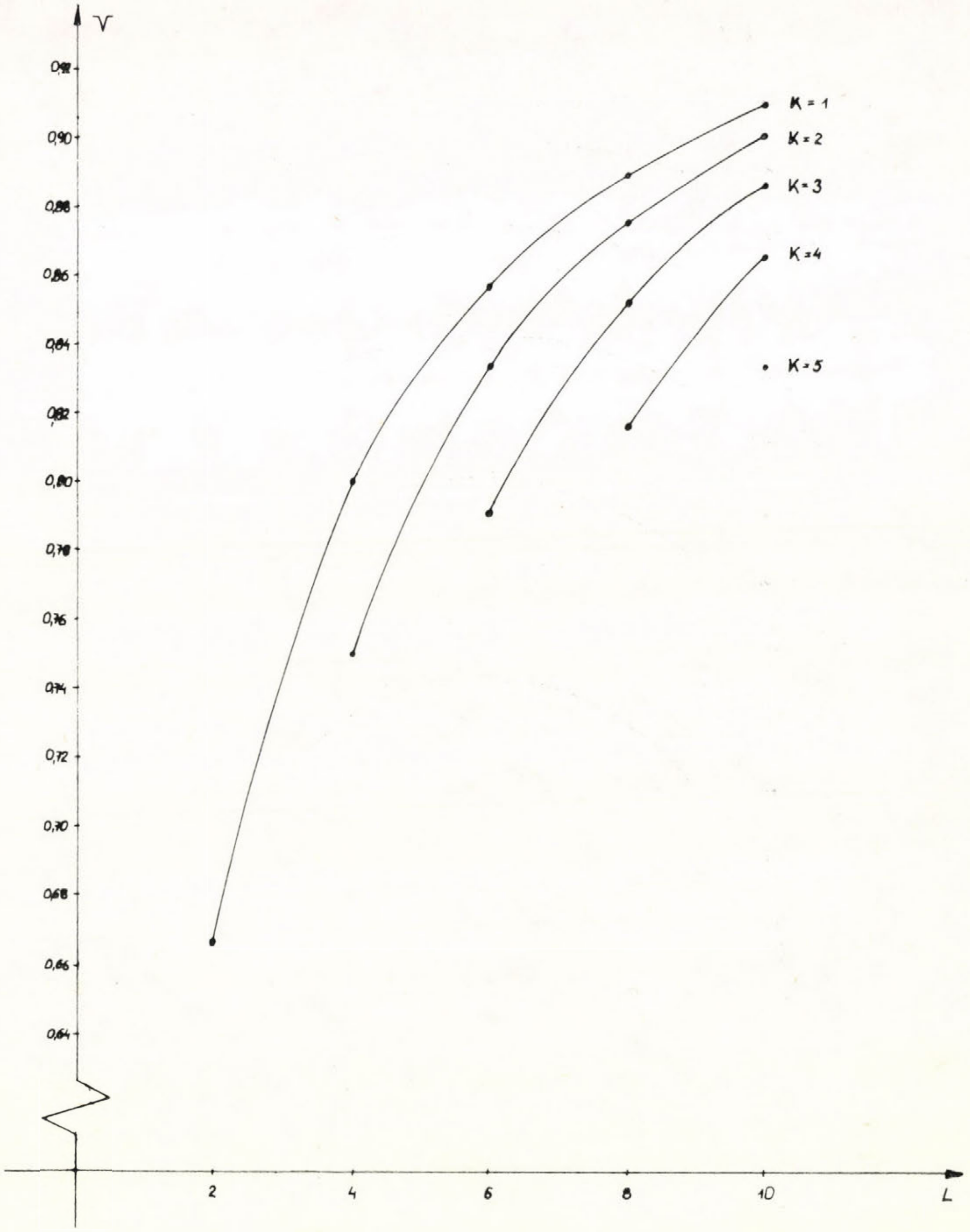

Fig. 6. Comparison of system capacity in the case of constant system accumulation L and different buffer size N as well as number of servers K, L = 2K + N. Data taken from [53].

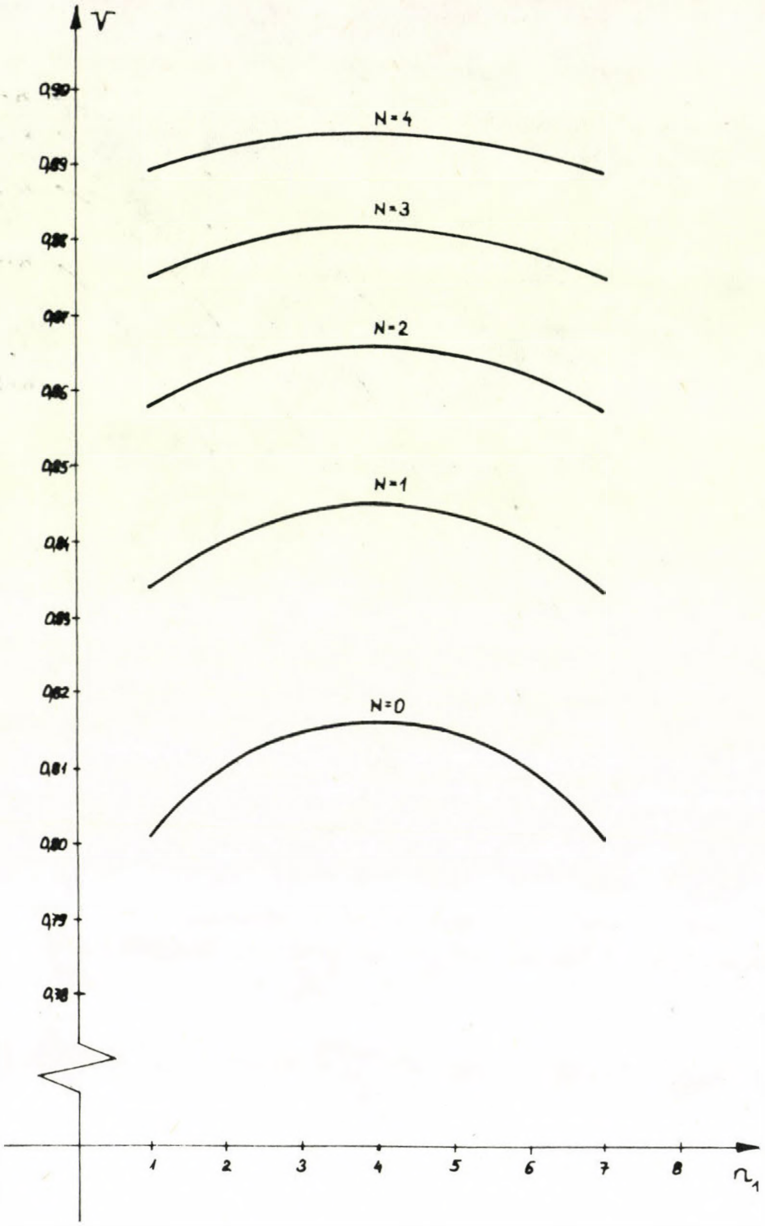

Fig. 7. Comparison of system capacity in the case of constant buffer size N and different number of servers at individual stages, $n_1 \neq n_2$, $n_1 + n_2 = const.$ Data taken from [53].

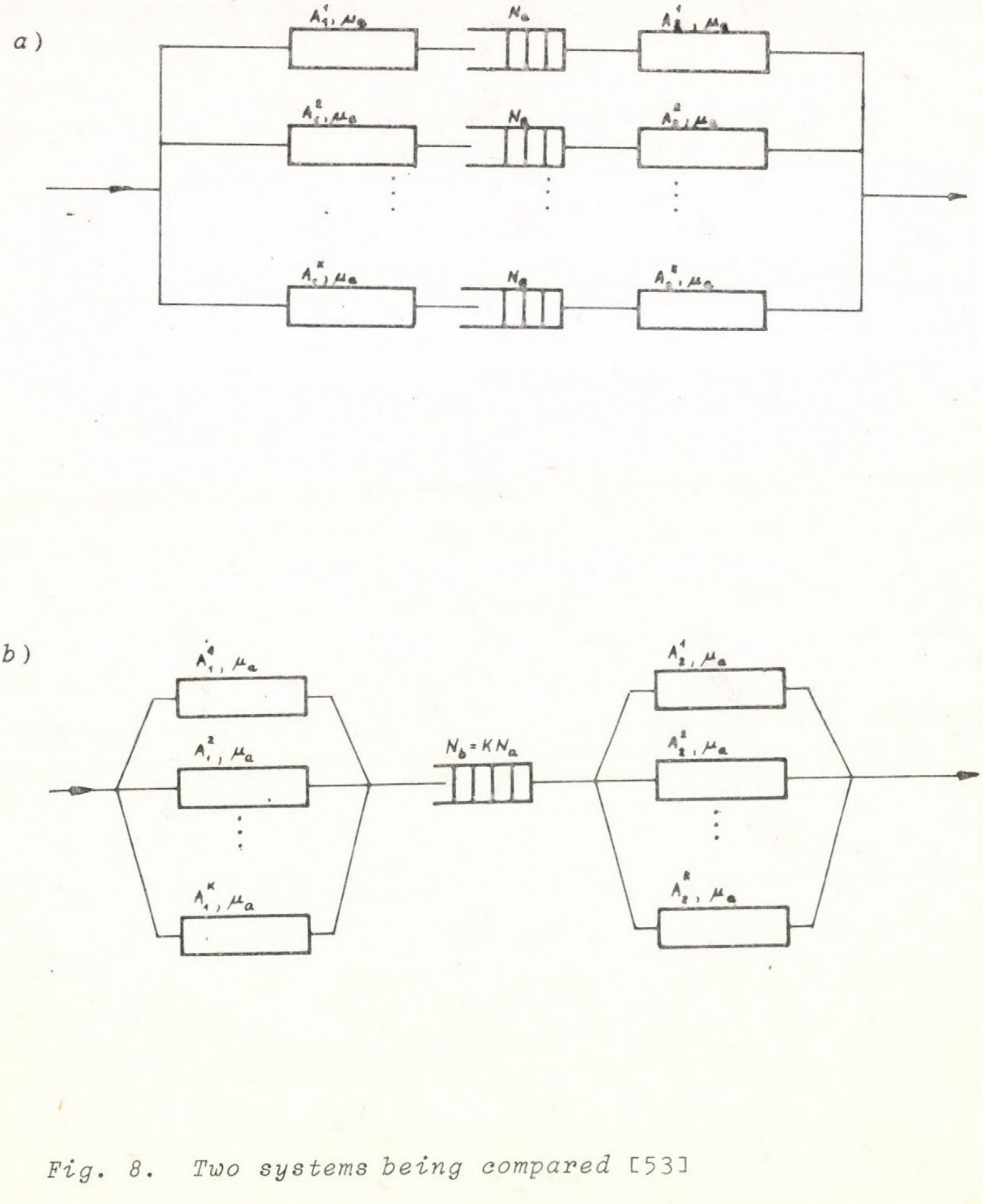

Fig. 8. Two systems being compared [53]

a) K queueing lines, each with intermediate buffer of size N_a

b) homogeneous two-stage queueing system with buffer size $N_b = k\ N_a$

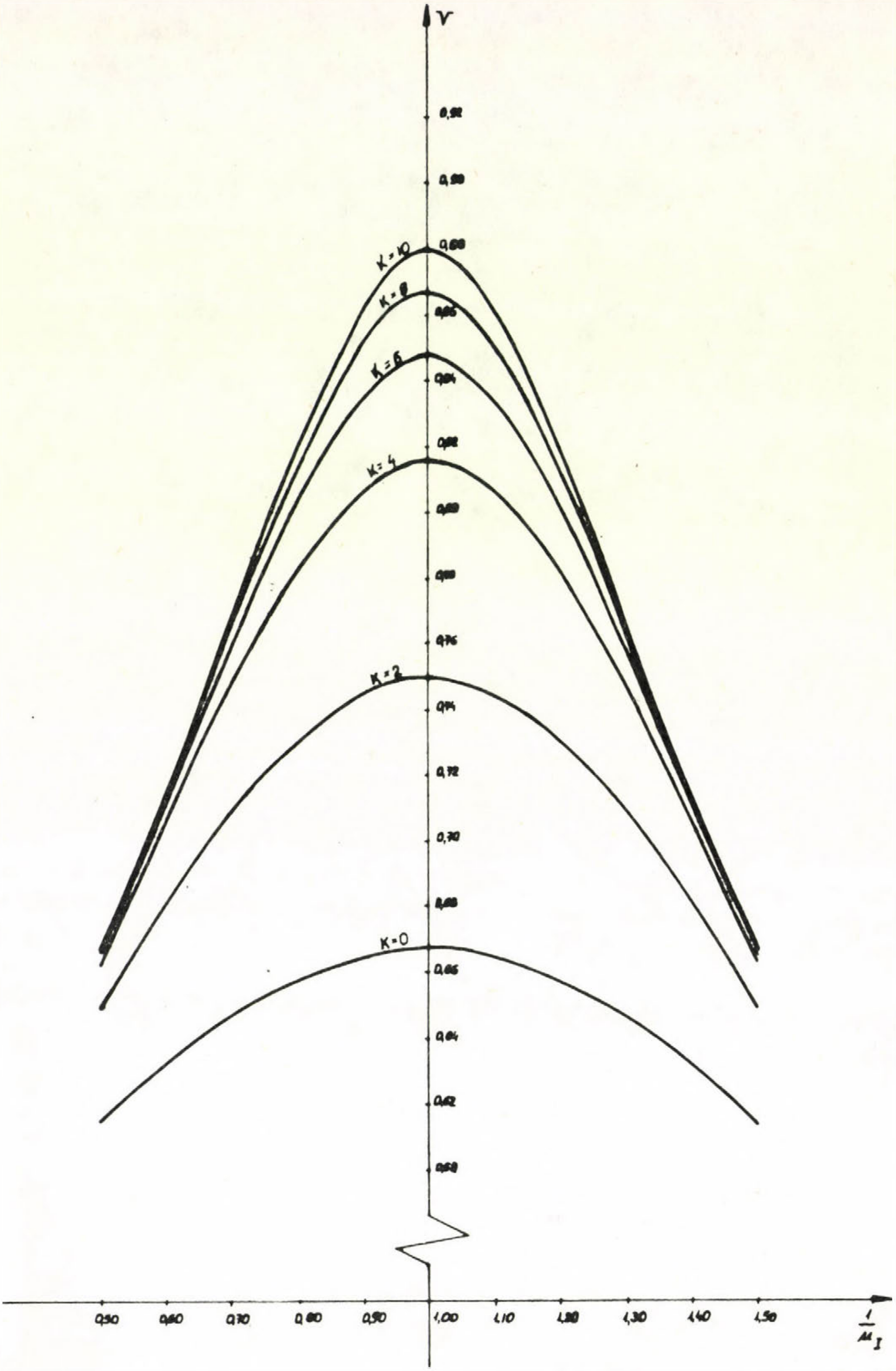

Fig. 9. The capacity of a two-stage unbalanced homogeneous system [53] having equal number of servers at both stages. N = 0.

Fig.10
Hol.52

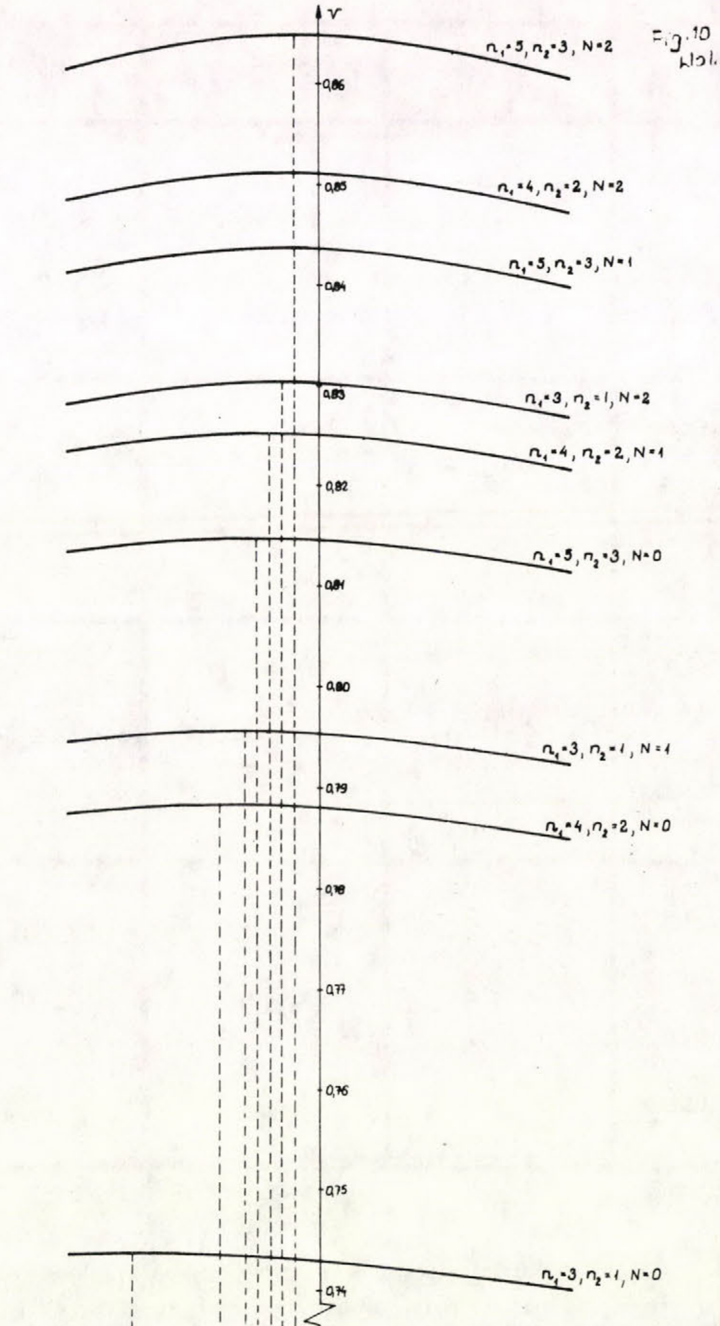

Fig. 10. The capacity of a two-stage unbalanced queueing system [53] with different number of servers at individual stages.

	M=3			M=4
	N=0	N=2	N=4	N=0
Optimal unbalancing $E(b_2)$	0.82 /0.83/	0.92	0.94	0.86
Throughput increase over the balanced case [%]	100.5 /100.54/	100.4	100.3	100.9
Range of unbalancing preserving maximal gain -0.1%	0.74-0.92	0.86-0.96	0.92-0.98	0.82-0.92
Unbalancing /in proper direction preserving the throughput of the balanced case	0.66	0.82	0.88	0.72

Table 1. The capacity of unbalanced three and four- stage queueing lines. Based upon data from [18] , data in parenthesis taken from [41]

Optimal unbalancing	$E(b_1) = 1.1$ $E(b_2) = 1.06$ $E(b_3) = 1.02$ $E(b_4) = 0.98$ $E(b_5) = 0.94$ $E(b_6) = 0.9$
Unbalancing /in proper/ direction preserving the throughput of the balanced case	$E(b_1) = 1.2$ $E(b_2) = 1.12$ $E(b_3) = 1.04$ $E(b_4) = 0.96$ $E(b_5) = 0.88$ $E(b_6) = 0.80$
Throughput increase over the balanced case [%]	1 %

Table 2. Some data characterizing the capacity of an unbalanced queueing line with $M=12$, $N=0$, according to [41] . An equality $E(b_i) = E(b_{12-i})$ holds.

The pattern considered	Capacity of balanced case	Gain from unbalancing	Optimal unbalancing		
			$E(b_1)$	$E(b_2)$	$E(b_3)$
a/ E E D	0.6160	102.80	1.00	0.73	1.27
b/ D E E	0.6160	102.80	1.27	0.73	1.00
c/ E D E	0.6167	100.35	0.945	1.110	0.945
d/ E D D	0.7311	106.79	0.62	1.19	1.19
e/ D D E	0.7311	106.79	1.19	1.19	0.62
f/ D E D	0.7311	106.79	1.19	0.62	1.19
g/ D D D	1	No	-----		
h/ E E E	0.5641	100.5	1.09	0.82	1.09

Table 3. Optimal unbalancing of three-stage systems with different service time distributions /based on [40] /. The last pattern utilizes data from [18]
E stands for exponential and D for deterministic service time distribution.

Service time distribution at the second stage						
C_2	Uniform			Erlang		
	$C_1=0$, N=0	$C_1=1$, N=0	$C_1=1$, N=1	$C_1=0$, N=0	$C_1=1$, N=0	$C_1=1$, N=1
0.00	1.0	0.73106	0.82366	1.0	0.73106	0.82366
0.10	0.95850	0.73008	0.82263	0.91670	0.73008	0.82263
0.20	0.92030	0.72712	0.81954	0.92633	0.72721	0.81960
0.30	0.88503	0.72220	0.81443	0.89372	0.72262	0.81471
0.40	0.85237	0.71530	0.80730	0.86530	0.71635	0.80773
0.50	0.82203	0.70640	0.79824	0.83655	0.70942	0.80029

Zable 4. Capacity of a two-stage, balanced queueing line with either regular /C_1 =0/ or exponential /C_1 =1/ service time distributions at the first stage and two different service time distributions with variability coefficien C_2 at the second stege. Data taken from [39] .

N	0	1	2	3	4	5	6	7	8	9	10
V	0.66	0.75	0.8	0.833	0.857	0.876	0.889	0.9	0.909	0.917	0.924
E	12.5	6.66	4.17	2.86	2.08	1.59	1.25	1.01	0.834	0.7	0.595

Table 5. Values of V and E versus N for the two-stage system with exponential servers.

		$N_a = 0$	$N_a = 1$	$N_a = 2$
K queueing lines		0.6667	0.7500	0.8000
Multistage Queueing Systems with multiservers	K = 2	0.7500	0.8333	0.8750
	K = 3	0.7900	0.8714	0.9072
	K = 4	0.8161	0.8940	0.9256

Table 6. Comparison of the capacity achieved by a MQS and corresponding set of queueing lines.Data taken from [53] .

Összefoglalás

Véges tárolókapacitású többlépcsős sorbanállási modellek átbocsátóképességének optimalizálása

A többlépcsős sorbanállási modellek mostanában nagy figyelemnek örvendenek, mert jól használhatók számos ipari rendszer modellezésére. Ha a különböző lépcsők kiszolgálási ideje eltérő, akkor közöttük sorok jöhetnek létre és a sorok hossza a gyakorlatban korlátozott. A sorhossz túllépése az igény elvesztését eredményezi. Ezt a jelenséget "blokkolvasással" kerülik el.

Sok cikk foglalkozott már a többlépcsős rendszerekkel, de még nem állnak rendelkezésre egzakt analitikus eszközök arra az esetre, ha a kiszolgálók száma háromnál több, és kisebb rendszerekben is csak egyes speciális eseteket vizsgáltak részletesen. A cikk célja az átbocsátóképesség-optimalizálás jelenlegi helyzetének bemutatása. A feladat formális felállítása után a cikk felvázolja az átbocsátóképességet befolyásoló egyes paraméterek optimális kiválasztását.

РЕЗЮМЕ

Об оптимизации пропускной способности многофазных систем массового обслуживания с конечной очередью в отдельных фазах

Многофазные системы массового обслуживания имеют большое значение, так как их хорошо можно использовать для моделирования различных промышленных систем. Несмотря на то, что внимание ряда математиков обратилось на проблемы систем, в настоящее время нет точных аналитических средств для исследования систем, в которых число обслуживающих приборов больше трех. Для систем, в которых это число не больше трех, только частные случаи рассматриваются подробно.

Целью данной работы является показ достижений в области оптимизации пропускной способности. После постановки задачи в статье подробно изучаются основные параметры, от которых зависит пропускная способность систем.