WSSIAA 5 (1995) pp. 1–16

Static Optimization of Queueing Systems

Onno J. Boxma
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands;
Tilburg University, Faculty of Economics
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

This paper discusses some recent developments in the static optimization of queueing systems. Special attention is given to three problem classes: (i) the optimal allocation of servers, or service capacity, to queues in a network; (ii) the optimal allocation of the visits of a single server to several queues (a polling system); (iii) the optimal allocation of a single arrival stream to several single server queues.

1. Introduction

When several users compete for the use of a common resource, the limited capacity of the resource can give rise to congestion. This situation occurs in a plethora of everyday situations: people queue at a counter in a bank or supermarket, congestion occurs in road traffic, products encounter delays at machines during their production process, messages wait for access to a common transmission channel and computer jobs for the use of a set of processors.

Queueing occurs even when the service capacity of the resource strongly exceeds the demand. This is due to the fact that the interarrival times of the users, and their required service times, are generally not fixed. A mathematical model of congestion phenomena therefore usually represents interarrival and service times of users by random variables. The resources are called service facilities, with a single server or multiple servers, and the users are called customers. Customers often visit a number of service facilities, encountering several queues during their stay in the system.

Queueing theory is devoted to the description, analysis and optimization of such *queueing systems*. It concentrates on a few key *performance measures*, like queue lengths and waiting times. Due to the stochastic nature of the arrival and service processes, and of the routing process of customers through a network of queues, the main performance measures are also random variables (or moments thereof). Generally, costs are in a natural way associated with these performance measures. The ultimate goal of performance analysis is optimization - and that is the subject of this paper.

While an enormous amount of literature has been devoted to the probabilistic analysis of queueing systems, their optimization is somewhat lagging behind. This is partly due

to the mathematical complexity of queueing systems: only rarely does one find nice structural properties or simple explicit expressions that allow straightforward optimization. Still, a sizable literature discusses the optimization and control of queues. One possible classification of this literature is according to the aspect of the queueing system to which it refers: (i) facility lay-out, (ii) admission control, (iii) customer routing, (iv) processing capacity ((re-)allocation of numbers of servers, and also control of the speed of servers), (v) service order, and (vi) buffer allocation. Some of these problems (facility lay-out, buffer allocation) typically occur in the design and fine-tuning phases of a service facility, whereas other problems mainly arise in its daily operation.
This brings us to another classification: *static* versus *dynamic* queueing system optimization. Consider for example a stream of jobs that have access to N parallel processors, possibly with different processing speeds. Suppose that the jobs must be allocated to the processors in a way that minimizes the mean waiting cost, different costs being assigned to one unit of delay at each of the processors. An important element in this customer assignment problem is the available information; this can range from 'complete observation', i.e., total knowledge about the system at any point in time (including exact queue lengths and service times), to only information about some basic characteristics like arrival rates or mean service times. In general, the term *dynamic* is used for policies which operate under time-dependent information, whereas policies operating under time-independent characteristics are called *static*.
Clearly, the more information is available for making decisions, the better the allocation can be. Dynamic policies in general perform better than static policies. However, static allocation policies are also of considerable interest. First of all, the situation of total knowledge at all times is unrealistic. From a viewpoint of costs, overhead grows as the amount of information to be exchanged, stored and processed increases. Furthermore, dynamic policies are not always that effective: there will always be some delay between updates of the system's current state, and this may have a considerable effect upon the quality of the policy. Moreover, it may be extremely difficult or time-consuming to solve a control problem under time-dependent information, while only rarely the structure of the optimal policy can be fully determined. Static policies, which often lend themselves more easily to performance analysis, can then be employed to provide performance indications (e.g., bounds) for dynamically controlled systems.

In this paper we restrict ourself to (a selection of) *static* queueing optimization problems. We refer to Stidham and Weber [29] for a survey of dynamic control problems in queueing networks, with an emphasis on models based on Markov decision theory; Chapter 8 of Walrand [32] is also highly recommended, for the structural insight it provides in the dynamic control of queues.
In Section 2 we discuss the optimal (re-)allocation of servers to queues in a network; we also pay attention to the assignment of service capacity, in the form of service speeds, to the single server queues of a network. Section 3 is devoted to the optimization of a *polling* system, i.e., a multiqueue system with only one server who moves from queue to queue. The service disciplines of the server at the various queues are studied, as well

as the optimal route of the server along the queues. Section 4 considers a problem that is in some sense dual to the latter problem: the earlier mentioned optimal allocation of customers to several queues in parallel. We always assume that buffer capacity is unlimited; for buffer allocation problems we refer to the survey [34].

Remark 1.1
Most of the optimization problems that are discussed in this paper have the following property: the objective function to be minimized by choosing a set of parameters $v_1, \ldots, v_N$ can be separated into N terms, the ith being a function of v_i only, that is convex in v_i, $i = 1, \ldots, N$. This is characteristic of a class of resource allocation problems discussed in the book of Ibaraki and Katoh [18]. They present several algorithms for such problems. The required convexity/concavity properties of the performance measures of queueing systems have only recently been studied systematically; see Liyanage and Shanthikumar [25] and Buzacott and Shanthikumar [11] and references therein. Important references are in particular a series of papers by Shaked, Shanthikumar and Yao that develop a sample-path based approach to obtain structural properties of queueing systems; see e.g. Shaked and Shanthikumar [28].

2. Static service capacity allocation

Consider an open Jackson network of M/M/. queues $Q_1, Q_2, \ldots, Q_N$. We discuss the following problems. (i) The server reallocation problem: how should a pool of M servers be distributed over the queues such as to minimize a weighted sum of the mean numbers of customers? (ii) The server allocation problem: how many servers should be allocated to each station, such that a weighted sum of the mean numbers of customers is below a certain level while minimizing costs? Or, dually: how many servers should be allocated to each station, such that server investment costs are kept below a certain level while minimizing a weighted sum of the mean numbers of customers? (iii) Kleinrock's capacity assignment problem: allocate server speeds to N single-server stations such that investment costs are kept below a certain level while minimizing the mean sojourn time of a customer in the network. The dual problem is also considered.

(i) *The server reallocation problem*
Let λ_i be the total arrival rate (external plus internal, the latter being determined by the Markovian routing matrix) at Q_i of the Jackson network, and let μ_i be the service rate of each server at Q_i. Hence $m_i^L := \lfloor \lambda_i/\mu_i \rfloor + 1$, with $\lfloor\ \rfloor$ the integer rounddown operation, denotes the minimal number of servers at Q_i such that the traffic intensity at Q_i is less than one. The server reallocation problem (**SR**) for the open Jackson network is formulated as follows:
SR

$$\mathrm{Min}_{m_1,\ldots,m_N} \sum_{i=1}^{N} c_i \mathrm{EL}_i(m_i) \tag{2.1}$$

$$s.t. \sum_{i=1}^{N} m_i = M, \quad m_i \geq m_i^L, \quad i = 1, \ldots, N.$$

Here $\mathrm{EL}_i(m_i)$ denotes the mean number of customers at Q_i when this queue has m_i servers, and c_i is a cost factor. It is proven in [9] that the following greedy, or marginal allocation, algorithm is optimal for the **SR** problem. Start by allocating m_i^L servers to Q_i, $i = 1, \ldots, N$. At each iteration step add one server to that queue where the greatest decrement in the objective function is achieved. Repeat this procedure until all M servers have been allocated. The optimality of this marginal allocation algorithm is typical for resource allocation problems in which the objective function is separable into convex terms while the constraints are *linear*.

(ii) *The server allocation problem*
Again consider the open Jackson network. The server allocation problem (**SA**) is formulated as follows:
SA

$$\mathrm{Min}_{m_1,\ldots,m_N} \sum_{i=1}^{N} F_i(m_i), \tag{2.2}$$

$$s.t. \sum_{i=1}^{N} c_i \mathrm{EL}_i(m_i) \leq W, \qquad m_i \geq m_i^L, \quad i = 1, \ldots, N.$$

Here $F_i(m_i)$ is a convex and decreasing function of m_i, that denotes the investment costs involved in allocating m_i servers to Q_i; W is a given number. If, e.g., $c_i \equiv 1/\mu_i$ then W indicates an upper bound on the mean total workload in the system.
Problem **SA** can be regarded as a generalization of the knapsack problem. Hence it is NP-complete. In [9] a simple greedy heuristic is proposed that represents a useful approach to the solution of problem **SA**: Start by allocating m_i^L servers to Q_i. At each iteration step, allocate one server to the queue for which the ratio of the increment of the objective function and the decrement of the weighted sum of mean queue lengths is the smallest. Stop as soon as adding a server makes the allocation feasible.
In [14] two algorithms are proposed that build upon this algorithm. They lead to substantially better results, at the expense of the complexity increasing by a factor N respectively N^2. For those two algorithms also worst-case performance ratios of 2 respectively 3/2 are proven in [14], whereas for the greedy heuristic it has only been proven that the minimal value of the objective function lies in between the values of the one-but-last (infeasible) and last allocations. Van Vliet and Rinnooy Kan [31] extend the greedy heuristic of [9] in another way. They allow general external interarrival time distributions and general service time distributions at the queues, and they use a two-moment parametric decomposition approach to estimate the first two moments of all interarrival times; subsequently they approximate the mean numbers of customers

in the queues, based on the first two moments of interarrival and service times at each queue in isolation. For the latter approximation they use a known approximation formula that is convex in the number of servers.
Aarts [1] considers the dual server allocation problem (**DSA**):
DSA

$$\text{Min}_{m_1,\ldots,m_N} \sum_{i=1}^{N} c_i \text{EL}_i(m_i), \tag{2.3}$$

$$s.t. \sum_{i=1}^{N} F_i(m_i) \leq F, \qquad m_i \geq m_i^L, \quad i = 1, \ldots, N,$$

with F some constant. He discusses the greedy heuristic and two refinements, similar to the three algorithms mentioned for (2.2). He observes that if $F_i(m_i) = dm_i$, $i = 1, \ldots, N$, then problem (2.3) amounts to the server reallocation problem (2.1), which is solved exactly by the greedy heuristic.

(iii) *The capacity assignment problem*
Consider an open Jackson network of M/M/1 queues, with service capacity μ_i at queue Q_i. Let λ_i denote the total (external plus internal) arrival rate at Q_i. The mean total sojourn time ET of an arbitrary customer in the network is given by

$$\text{ET} = \frac{1}{\gamma} \sum_{i=1}^{N} \frac{\mu_i}{\lambda_i - \mu_i}.$$

Here γ denotes the total external arrival rate in the network. Kleinrock [20] has posed and solved the following capacity assignment problem (**CA**); he has formulated it in the framework of assigning channel capacities in a communication network.
CA

$$\text{Min}_{\mu_1,\ldots,\mu_N} \text{ET}, \tag{2.4}$$

$$s.t. \sum_{i=1}^{N} d_i \mu_i \leq D, \quad \mu_i > \lambda_i, \quad i = 1, \ldots, N.$$

Again the separability/convexity structure appears. This time the problem even allows an explicit solution, that is easily obtained using the Lagrange multiplier technique. The optimal service rates turn out to be [20]:

$$\mu_i^* = \lambda_i + \frac{D - \sum_{j=1}^{N} \lambda_j d_j}{d_i} \frac{\sqrt{\lambda_i d_i}}{\sum_{j=1}^{N} \sqrt{\lambda_j d_j}}, \quad i = 1, \ldots, N. \tag{2.5}$$

After having allocated λ_i service capacity to Q_i, the remaining funds are invested proportionally to the square root of λ_i and d_i.

Remark 2.1
The dual capacity assignment problem (**DCA**) reads:
DCA

$$\text{Min}_{\mu_1,\ldots,\mu_N} \sum_{i=1}^{N} d_i\mu_i, \tag{2.6}$$

$$s.t.\ \text{ET} \leq T, \quad \mu_i > \lambda_i, \quad i = 1,\ldots,N,$$

with T some positive constant. It is solved in exactly the same way, yielding a very similarly structured solution. One easily verifies [1] that, if T equals the optimal value in problem **CA**, then the optimal value of $\sum_{i=1}^{N} d_i\mu_i$ in the dual problem equals D.

Remark 2.2
Problems **CA** and **DCA** have been extended in several directions. We refer to Bitran and Tirupati [2] for a heuristic approach (related to those for **SA**) to **DCA** in the case of general service time distributions.

3. Static polling optimization

The standard polling system is a queueing system in which a single server, S, visits N queues $Q_1,\ldots,Q_N$ in some order. Polling systems presently receive much attention, partly because of their ability to model many resource allocation phenomena in computer-communications. After a brief model description, we discuss: (i) optimal server routing and (ii) optimal service behaviour at the queues.

Model description
S serves N infinite-capacity queues $Q_1,\ldots,Q_N$. Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is λ_i, $i = 1,\ldots,N$. Customers arriving at Q_i are called class-i customers; their service times are independent random variables $\mathbf{B}_i$ with mean β_i and second moment $\beta_i^{(2)}$, $i = 1,\ldots,N$. After their service at Q_i they leave the system. The offered traffic load, ρ_i, at Q_i is defined as $\rho_i := \lambda_i\beta_i$, $i = 1,\ldots,N$, and the total offered load is called ρ. When swapping out of Q_i, the server incurs a switchover period of type i; the switchover durations of type i are independent random variables $\mathbf{S}_i$ with mean s_i and second moment $s_i^{(2)}$, $i = 1,\ldots,N$. All interarrival, service and switchover processes are independent stochastic processes. The *service discipline* at Q_i determines how many customers are served when S visits Q_i. Important disciplines are:

- Exhaustive (E): S serves Q_i until it has become empty.
- Gated (G): S serves exactly the customers that were present at the beginning of the visit.

- k-limited ($k-L$): S serves k customers or empties Q_i, whichever happens first.

Furthermore, for some applications (traffic lights, timed-token protocol in a local area network), the visit period is restricted by a fixed time limit. In the application of a signalized traffic intersection, it may be natural to leave the server (green light) at a queue even if it is empty; but in the now following discussion of polling optimization we shall restrict ourself to the case in which S never resides at an empty queue. In the sequel we assume the traffic parameters to be such that the polling system is ergodic. A necessary condition for this to hold is $\rho < 1$; it is known to be also sufficient when the service discipline at all queues is E or G, but not when, e.g., a queue is $k-L$.

(i) *The optimal server routing problem*
We start with a simple result for a *probabilistic polling model*, viz., a polling model in which S visits the queues according to a probabilistic routing mechanism: with probability p_i it chooses Q_i for its next visit, $i = 1, \ldots, N$. Suppose that all switchover times between the various queues have mean σ and second moment $\sigma^{(2)}$, and that service at a set of queues e is exhaustive while it is gated at the remaining set, g.
Let EW_i denote the mean waiting time at Q_i, $i = 1, \ldots, N$. Consider the following optimal server routing problem (**OSR**):
OSR

$$\mathrm{Min}_{p_1,\ldots,p_N} \sum_{i=1}^{N} \rho_i \mathrm{EW}_i \tag{3.1}$$

$$s.t. \quad \sum_{i=1}^{N} p_i = 1, \quad p_1 \geq 0, \ldots, p_N \geq 0.$$

It is easily seen that minimizing this objective function is equivalent to minimizing the mean *workload*. In this model

$$\sum_{i=1}^{N} \rho_i \mathrm{EW}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} - \frac{\sigma}{1-\rho} \sum_{k \in e} \frac{\rho_k^2}{p_k} + \frac{\sigma}{1-\rho} \sum_{k=1}^{N} \frac{\rho_k}{p_k} - \rho\sigma + \rho \frac{\sigma^{(2)}}{\sigma}. \tag{3.2}$$

We now have a classical non-linear optimization problem with linear constraints. Using standard Lagrange multiplier techniques we find the following solution (cf. [7]):

$$k \in e: \quad p_k = \frac{\sqrt{\rho_k(1-\rho_k)}}{\sum_{j \in e} \sqrt{\rho_j(1-\rho_j)} + \sum_{j \in g} \sqrt{\rho_j}}; \tag{3.3}$$

$$k \in g: \quad p_k = \frac{\sqrt{\rho_k}}{\sum_{j \in e} \sqrt{\rho_j(1-\rho_j)} + \sum_{j \in g} \sqrt{\rho_j}}. \tag{3.4}$$

We refer to [27] for an extensive study of the optimization of *Markovian* server routing (i.e., server routing probability p_{ij} from Q_i to Q_j). Below we turn to the following related problem, again for a mixture of E and G queues and i.i.d. switchover times. S is allowed to visit the queues according to a fixed pattern, a *polling table* (like $Q_1Q_2Q_3 \ldots Q_N$, which is called cyclic polling, or $Q_1Q_2Q_1Q_3Q_1 \ldots Q_1Q_N$, which is called star polling). We seek the polling table that minimizes the mean workload, or equivalently, $\sum_{i=1}^{N} \rho_i \mathrm{EW}_i$. An implicit expression for this sum is known, but it does not allow a straightforward optimization. If an upper bound on the table size is given, then an integer programming problem results; but without such a restriction it is not clear a priori whether a given 'good' table cannot be improved by taking a much larger table with a very similar structure. In [7] an approximate approach is proposed to the problem of choosing an optimal polling table, and shown to perform well. It consists of three steps.

Step 1. Choose the occurrence frequencies of the queues in the table by taking the optimal frequencies p_k obtained in (3.3) and (3.4) for the probabilistic polling model with the same traffic parameters.
Step 2. Based on those occurrence frequencies, determine a 'good' table size M (take the smallest possible M such that for all k, Mp_k is within a predetermined small distance from a positive integer).
Step 3. Given the round-off occurrence numbers, say n_i, find a table such that Q_i occurs n_i times with the visits to each particular queue as evenly spaced as possible. This spacing problem is handled using the so-called *Golden Ratio* policy, cf. Itai and Rosberg [19].

Remark 3.1
Comparison of the optimal mean workloads in the cases of probabilistic polling and polling tables reveals that the latter minimum is considerably lower. The explanation is that the variance of the time between successive visits to a particular queue is much smaller in a polling table; this more regular behaviour leads to a smaller mean workload. Still, the occurrence frequencies of the queues in the optimal polling table are well predicted by the occurrence frequencies in the probabilistic polling model.

Remark 3.2
The Golden Ratio policy has recently been applied to several 'even spacing' problems. In [5], in the context of optimizing the *time limits* in a polling model, an algorithm is proposed which seems to outperform Golden Ratio; it is based on a neat result of Hajek [16] concerning the optimal splitting of point processes.

Let us now consider the somewhat more general polling table optimization problem with objective function:

$$\text{Min} \sum_{i=1}^{N} c_i \lambda_i \mathrm{EW}_i. \tag{3.5}$$

Here c_i are nonnegative constants that reflect the cost of waiting one unit of time at Q_i. The choice $c_i \equiv \beta_i$ again yields the objective function of (3.1), while $c_i \equiv 1$ leads to the minimization of the mean total number of waiting customers in the system (note that Little's formula implies that $\lambda_i \mathbf{EW}_i$ equals the mean number of waiting customers at Q_i). Let us also allow $1-L$ service at one or more of the queues, and non-identically distributed switchover times. Even probabilistic polling no longer yields an explicit analytic solution when one of these changes is introduced. For the corresponding (and practically more important) polling table minimization problem, step 1 above now has to be adapted after which steps 2 and 3 can again be used. In [8] some approaches are developed for this adaptation. One of them is based on the following approximations for $\mathbf{EW}_i$, for a polling table with n_i evenly spaced occurrences of Q_i:

$$i \in e: \quad \mathbf{EW}_i \approx A(1-\rho_i)\frac{\sum_{j=1}^{N} n_j s_j}{n_i}; \tag{3.6}$$

$$i \in g: \quad \mathbf{EW}_i \approx A(1+\rho_i)\frac{\sum_{j=1}^{N} n_j s_j}{n_i}; \tag{3.7}$$

$$i \in 1-L: \quad \mathbf{EW}_i \approx A\frac{1-\rho+\rho_i}{1-\rho-\lambda_i \sum_{j=1}^{N} n_j s_j/n_i}\frac{\sum_{j=1}^{N} n_j s_j}{n_i}. \tag{3.8}$$

A denotes some positive constant which we don't need to specify, as only the ratio of the various mean waiting times matters for the optimization. Minimization of $\sum_{i=1}^{N} c_i \lambda_i \mathbf{EW}_i$ easily yields the optimal n_i values, up to a multiplicative constant:

$$i \in e: \quad n_i \sim \sqrt{c_i \lambda_i (1-\rho_i)/s_i}; \tag{3.9}$$

$$i \in g: \quad n_i \sim \sqrt{c_i \lambda_i (1+\rho_i)/s_i}; \tag{3.10}$$

$$i \in 1-L: \quad n_i \sim \lambda_i + (1-\rho-\sum_{k=1}^{N} \lambda_k s_k)\frac{\sqrt{c_i \lambda_i (1-\rho+\rho_i)/s_i}}{\sum_{j=1}^{N} s_j \sqrt{c_j \lambda_j (1-\rho+\rho_j)/s_j}}. \tag{3.11}$$

If all queues are $1-L$ and we add the constraint $\sum_{j=1}^{N} n_j s_j/(1-\rho) = C^*$, which amounts to prescribing that the mean cycle time of the polling table equals C^*, then the resulting 1-limited polling table problem (**LPT**) is:

LPT

$$\mathrm{Min}_{n_1,\ldots,n_N} \sum_{i=1}^{N} \frac{c_i \lambda_i (1-\rho+\rho_i)}{n_i - \lambda_i C^*} \tag{3.12}$$

$$s.t. \sum_{j=1}^{N} n_j s_j/(1-\rho) = C^*.$$

It has a similar form as problem **CA** of Section 2. The solution is given by (3.11) when its righthand side is multiplied by C^*. Its interpretation is that Q_i should be visited at least $\lambda_i C^*$ times per cycle, as this is the mean number of arrivals at Q_i during a cycle of mean C^*; the remaining 'capacity' is allocated according to a square root rule. In [8] the resulting rule is shown to perform quite satisfactorily.

Remark 3.3
An interesting application of polling table optimization occurs in stochastic economic lot scheduling, where several items need to be produced in a common facility. A. Federgruen (personal communication) studies the problem of determining an optimal production sequence (with a somewhat more involved objective function), using a three-step procedure as outlined above.

Remark 3.4
We refer to Yechiali [35] for an interesting overview of semi-dynamic control of the server route in a polling system. Yechiali (see also Browne and Yechiali [10]) considers one-cycle look-ahead policies. For various service disciplines, he determines the visit order of the queues in the next cycle - in which all queues are visited exactly once - that minimizes the expected duration of that cycle.

(ii) *Optimal service behaviour at the queues*
Given a route of S along the queues, the question arises which service discipline should be used at each queue. Borst et al. [6] consider a cyclic polling model, viz., a polling model in which S visits the queues in cyclic order. The service discipline at Q_i is k_i-limited, $i = 1, \ldots, N$. They consider the following unconstrained k-limited problem (**UkL**) for this cyclic polling model:
UkL

$$\text{Min}_{k_1,\ldots,k_N} \sum_{i=1}^{N} c_i \lambda_i \text{EW}_i. \tag{3.13}$$

Next to this *unconstrained* optimization problem, they also consider the *constrained* k-limited problem (**CkL**) where the objective function in (3.13) must be minimized under the additional condition that $\sum_{i=1}^{N} \gamma_i k_i \leq K$ for some nonnegative parameters γ_i and constant K. Such a condition could reflect a limit on the (expected) cycle time of the server, which is relevant in some access protocols in local area networks.
Problem **CkL** is tackled in two different ways (both necessarily approximative: an exact analysis of polling models with k-limited service seems to be prohibitive in all but a few exceptional cases). One approach is to take such k_i values that the ratios $k_i : k_j$ agree with the ratios of the visit frequencies for a corresponding polling table with 1-limited service at all queues (see (3.11)). A minor adaptation of (3.8) has to be

made to incorporate the fact that switchover times out of Q_i occur once instead of k_i times. That results in the following mean waiting time approximation, with D_1 some constant and $s := \sum_{i=1}^{N} s_i$:

$$\mathrm{EW}_i \simeq D_1 \frac{1-\rho+\rho_i}{k_i(1-\rho)-\lambda_i s}. \tag{3.14}$$

A second approach is to use the Fuhrmann-Wang approximation [15] for EW_i in the cyclic polling model with k-limited service:

$$\mathrm{EW}_i \simeq D_2 \frac{(1-\rho_i)(1-\rho)+(2-\rho)\rho_i/k_i}{1-\rho-\lambda_i s/k_i}; \tag{3.15}$$

here D_2 is some constant. There is no need to specify the constants D_1 and D_2, as only the ratio of the mean waiting times plays a role in the optimization. Taking into account the constraint $\sum_{i=1}^{N} \gamma_i k_i \leq K$, the optimal solution k_i^*, $i = 1, \ldots, N$ of (3.13) using either (3.14) or (3.15) is found in a straightforward way. We mention only the one based on (3.15), as it is slightly better:

$$k_i^* = \frac{\lambda_i s}{1-\rho} + (K - \sum_{j=1}^{N} \gamma_j \frac{\lambda_j s}{1-\rho}) \frac{\sqrt{c_i \lambda_i [\rho_i(2-\rho)+\lambda_i s(1-\rho_i)]/\gamma_i}}{\sum_{j=1}^{N} \gamma_j \sqrt{c_j \lambda_j [\rho_j(2-\rho)+\lambda_j s(1-\rho_j)]/\gamma_j}}. \tag{3.16}$$

Again the resemblance with problem **CA** of Section 2 should be noted.
In [6] slightly better results for **CkL** are obtained by sharpening (3.15) somewhat, at the expense of no longer obtaining explicit expressions for the k_i; that improved approximation also yields good results for **UkL**.

Remark 3.5
An unproven conjecture in [6], that is used in the optimization procedure, is that EW_i is decreasing in k_i and nondecreasing in k_j for all $j \neq i$. Using a sample-path argument, Levy et al. [22] show that the service discipline that minimizes the mean workload, or equivalently $\sum_{i=1}^{N} \rho_i \mathrm{EW}_i$, in the unconstrained case is to serve all queues exhaustively (i.e., $k_i \equiv \infty$). Liu et al. [23], also allowing idling policies, in more generality discuss the server behaviour that stochastically minimizes the unfinished work and the number of customers in the system.

Remark 3.6
In the so-called Bernoulli service discipline, S serves after each service completion yet another customer with probability q_i and leaves with probability $1-q_i$. This discipline can be viewed as the stochastic counterpart of the k-limited discipline. Analogous to **UkL**, Blanc and Van der Mei [3] try to find those q_i that minimize the objective function (3.5). Their main approach is a numerical one, based on the use of the so-called power series algorithm.

Remark 3.7
A pioneering polling optimization study is due to Klimov [21]. It combines elements of both problem types studied in this section. Klimov studies a polling model without switchover times, but with probabilistic customer routing along the queues. After each service completion a decision is made as to which queue S should visit next. Klimov shows that the optimal server routing policy has an index structure: the optimal policy corresponds to an ordering of the queues such that S always chooses to serve a customer from the first nonempty queue in that order.

4. STATIC TRAFFIC ALLOCATION

Consider the following situation. Customers arrive according to a Poisson process with rate Λ. At the instance of arrival, a customer has to be assigned to one of N parallel single servers $Q_1, \ldots, Q_N$. The service time of a customer that is assigned to Q_i has distribution $B_i(\cdot)$ with mean β_i and second moment $\beta_i^{(2)}$, $i = 1, \ldots, N$. All service times are independent. Let P denote an allocation policy. Our aim is to minimize

$$\text{Min}_P \sum_{i=1}^{N} c_i f_i(P) \text{EW}_i(P). \tag{4.1}$$

The notation is as before, with now (P) indicating a dependency on the allocation policy P; the factors $f_i(P)$ are load-dependent weight factors. For example, if under P a fraction p_i of the customers is assigned to Q_i, then $f_i(P) = \Lambda p_i$ yields the objective function in (3.5).
We refer to Wang and Morris [33] for a survey on load balancing, including numerical comparisons of various dynamic and static allocation policies. We restrict ourself again to static policies. Buzen and Chen [12] have studied the *probabilistic* allocation, in which a customer is sent to Q_i with probability p_i, $i = 1, \ldots, N$. The arrival process at each queue now is a Poisson process, and hence each queue is an M/G/1 queue. They have used mathematical programming techniques to minimize the mean *sojourn* time of a customer. Let us also consider the class of probabilistic allocation policies, but with objective function (4.1). First take $f_i \equiv p_i$, and put $\lambda_i := \Lambda p_i$. The probabilistic allocation problem (**PA**) now reduces to:
PA

$$\text{Min}_{\lambda_1,\ldots,\lambda_N} \sum_{i=1}^{N} c_i \lambda_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\lambda_i \beta_i)} \tag{4.2}$$

$$s.t. \sum_{i=1}^{N} \lambda_i = \Lambda, \quad 0 \le \lambda_i \le \frac{1}{\beta_i}, \quad i = 1, \ldots, N.$$

It can also be verified that this optimization problem has a feasible solution provided that $\sum_{i=1}^{N} 1/\beta_i > \Lambda$, i.e., the arrival rate does not exceed the total service capacity.

Here and in the remainder of the section that is assumed to be the case.
Note that the objective function in (4.2) is separable into N terms, the ith term being strictly convex in λ_i; cf. Remark 1.1, where the reader is referred to the book of Ibaraki and Katoh [18]. Just like **CA**, this case is so simple that it can be solved explicitly, using Lagrange multiplier techniques; we find that the unique optimal rates λ_i^* are (cf. [13]):

$$\lambda_i^* = \frac{1}{\beta_i} - \frac{1}{\beta_i}[\sqrt{1 + \frac{2\beta_i\delta}{c_i\beta_i^{(2)}}}]^{-1}, \quad i = 1, \ldots, N, \tag{4.3}$$

in which the Lagrange multiplier δ is determined by the constraint $\sum_{i=1}^{N} \lambda_i = \Lambda$. If in (4.1) $f_i \equiv 1$ instead of $f_i \equiv p_i$, then problem **PA** has the same separable structure, with the control variables only interacting through the linear restriction $\sum_{i=1}^{N} \lambda_i = \Lambda$. In this case an explicit analytic solution is not obtained (in fact some of the λ_i^* now are equal to zero), but one can easily solve the problem numerically, using for example the algorithm RANK in [18], p. 19.
Algorithm RANK strongly depends on the strict convexity of the N terms. The convexity property implies that there is only one local minimum, which consequently has to be the optimal solution for the allocation problem. One of the cases in which the property of strict convexity may not hold is the probabilistic traffic allocation problem with a *general* arrival process, as studied in Tang and Van Vliet [30]. Their method involves the Frank-Wolfe algorithm, which was originally developed for quadratic programming and which provides a local minimum; they claim that it should be close to the global minimum.

Just like in polling optimization, in customer allocation one may expect to improve considerably upon a probabilistic allocation by allocating according to a fixed pattern. This expectation is based on the reduced variability of the arrival processes at the queues. In the sequel we assume that allocation is being done according to a fixed pattern, and we consider the problem of finding a 'good' pattern (w.r.t. the objective function (4.1)). Only in exceptional cases optimality of a pattern can be proven. We refer to Liu and Towsley [24] for such a result and further references. Liu and Towsley [24] consider the case of identical service time distributions at all queues, with an increasing failure rate. They show that the round-robin policy (send customers consecutively to $Q_1, Q_2, \ldots, Q_N, Q_1, \ldots$) minimizes, in the sense of a separable increasing convex ordering, the customer sojourn times and the numbers of customers in the queues.
Let $(a_1, a_2, \ldots, a_M)$ denote an allocation table, a_i indicating the number of the queue to which every $(i + kM)$-th customer is being sent, $k = 1, 2, \ldots$; $i = 1, \ldots, M$. Combé and Boxma [13] observe that the resulting arrival processes at the queues fall into the class of Markovian Arrival Processes (MAP), cf. [26]. Subsequently they present a three-step algorithm, similar to the three-step polling optimization algorithm of Section 3, to construct a 'good' pattern. The occurrence frequencies of the queues in

the allocation table are estimated in step 1, after which a suitable table size M and an even spacing are determined. It turns out that the optimal probabilistic allocation again gives a reasonable indication for the occurrence frequencies. However, the optimal probabilistic allocation seems to underestimate the traffic assignment to the queues with relatively high mean service time. The explanation is that the effect of 'regularizing' the arrival stream in pattern allocation is strongest for the queues with relatively large service times, so relatively small assignment probabilities. E.g., if the optimal probabilistic allocation fractions in a two-queue case are 8/9 and 1/9, then Q_2 would receive an Erlang-9 arrival process under pattern allocation, and Q_1 something close to Poisson (far less regular than Erlang-9). In step 1, instead of using the optimal probabilistic allocation, it is better to approximate the interarrival time distributions to the queues by Gamma distributions, subsequently approximate the mean waiting time in a Gamma/G/1 queue in a suitable way (cf. [13]) and solve the resulting non-linear optimization problem. That problem again has the separability/convexity structure as found in (4.2), and can be easily solved numerically (cf. [13]).

Remark 4.1
The approach using MAP and Gamma approximations seems applicable to several generalizations of the problem described above, like: the original arrival process is not a Poisson process; some of the queues also receive a 'dedicated' arrival stream; the arrival process must be allocated to *multiserver* queues in parallel.

Remark 4.2
Hordijk, Koole and Loeve [17] study the class of pattern allocation policies, for the special case of exponentially distributed service times, by using Markov Decision theory. Their algorithm leads to results of similar quality as those of [13].

Remark 4.3
Borst [4] considers the probabilistic allocation of not one but *several heterogeneous customer classes* to a set of parallel servers with different speeds. Each customer class has a Poisson arrival process and generally distributed service requirements. He studies the minimization of a weighted sum of the mean waiting times, exposing the structure of the optimal allocation. [11] considers a version of this problem in which all servers have the same speed, and in which the overall mean waiting time is the objective function. It would be interesting to study allocation patterns for this type of model.

Acknowledgement
The author's research has been supported by the European Grant BRA-QMIPS of CEC DG XIII.

REFERENCES

1. M.A.J. Aarts (1992). Production logistics - A queueing theoretic approach, *M.Sc. Thesis, Tilburg University*, in Dutch.
2. G.R. Bitran and D. Tirupati (1989). Trade-off curves, targeting, and balancing in

manufacturing networks, *Operations Research* **37**, 547-564.

3. J.P.C. Blanc and R.D. van der Mei (1992). Optimization of polling systems with Bernoulli schedules, *Report FEW 563, Tilburg University*, to appear in Performance Evaluation (1995).
4. S.C. Borst (1994). Optimal probabilistic allocation of customer types to servers, *CWI Report BS-R9415, Amsterdam*; to appear in the Proceedings of ACM Sigmetrics/Performance '95.
5. S.C. Borst, O.J. Boxma, J.H.A. Harink and G.B. Huitema (1994). Optimization of fixed time polling schemes, *Telecommunication Systems* **3**, 31-59.
6. S.C. Borst, O.J. Boxma and H. Levy (1993). The use of service limits for efficient operation of multi-station single-medium communication systems, *CWI Report BS-R9312, Amsterdam.*
7. O.J. Boxma, H. Levy and J.A. Weststrate (1990). Optimization of polling systems, In: *Performance '90*, eds. P.J.B. King, I. Mitrani and R.J. Pooley (North-Holland, Amsterdam) pp. 349-361.
8. O.J. Boxma, H. Levy and J.A. Weststrate (1991). Efficient visit frequencies for polling tables: minimization of waiting cost, *Queueing Systems* **9**, 133-162.
9. O.J. Boxma, A.H.G. Rinnooy Kan and M. van Vliet (1990). Machine allocation problems in manufacturing networks, *Eur. J. Oper. Res.* **45**, 47-54.
10. S. Browne and U. Yechiali (1989). Dynamic priority rules for cyclic-type queues, *Adv. Appl. Probab.* **21**, 432-450.
11. J.A. Buzacott and J.G. Shanthikumar (1992). Design of manufacturing systems using queueing models, *Queueing Systems* **12**, 135-213.
12. J.P. Buzen and P.P.-S. Chen (1974). Optimal load balancing in memory hierarchies, in: *Proc. IFIP 1974*, ed. J. Rosenfeld (North-Holland, Amsterdam) pp. 271-275.
13. M.B. Combé and O.J. Boxma (1994). Optimization of static traffic allocation policies, *Theoretical Computer Science* **125**, 17-43.
14. H. Frenk, M. Labbé, M. van Vliet and S. Zhang (1994). Improved algorithms for machine allocation in manufacturing systems, *Operations Research* **42**, 523-530.
15. S.W. Fuhrmann and Y.T. Wang (1988). Analysis of cyclic service systems with limited service: bounds and approximations, *Performance Evaluation* **9**, 35-54.
16. B. Hajek (1985). Extremal splittings of point processes, *Math. Oper. Res.* **10**, 543-556.
17. A. Hordijk, G.M. Koole, J.A. Loeve (1994). Analysis of a customer assignment model with no state information, *Prob. in the Eng. and Inform. Sciences* **8**, 419-429.
18. T.I. Ibaraki and N. Katoh (1988). *Resource Allocation Problems.* MIT Press, Cam-

bridge.

19. A. Itai and Z. Rosberg (1984). A Golden Ratio control policy for a multiple-access channel, *IEEE Trans. Autom. Control* **29**, 712-718.
20. L. Kleinrock (1964). *Communication Nets - Stochastic Message Flow and Delay.* McGraw-Hill, New York, 1964 (republished by Dover, 1972).
21. G.P. Klimov (1974). Time-sharing service systems I, *Theory of Prob. and its Appl.* **19**, 532-551.
22. H. Levy, M. Sidi and O.J. Boxma (1990). Dominance relations in polling systems, *Queueing Systems* **6**, 155-171.
23. Z. Liu, P. Nain and D. Towsley (1992). On optimal polling policies, *Queueing Systems* **11**, 59-83.
24. Z. Liu and D. Towsley (1994). Optimality of the round-robin routing policy, *J. Appl. Probab.* **31**, 466-475.
25. L. Liyanage and J.G. Shanthikumar (1992). Second-order properties of single-stage queueing systems, in: *Queueing and Related Models*, eds. U.N. Bhat and I.V. Basawa (Oxford Univ. Press, Oxford), pp. 129-160.
26. D.M. Lucantoni (1991). New results on the single server queue with a batch Markovian arrival process, *Stochastic Models* **7**, 1-46.
27. R.D. van der Mei (1995). Polling systems with Markovian server routing, *Report CentER, Tilburg.*
28. M. Shaked and J.G. Shanthikumar (1988). Stochastic convexity and its applications, *Adv. Appl. Probab.* **20**, 427-446.
29. S. Stidham, Jr. and R. Weber (1993). A survey of Markov decision models for control of networks of queues, *Queueing Systems* **13**, 291-314.
30. C.S. Tang and M. van Vliet (1994). Traffic allocation for manufacturing systems, *Eur. J. Oper. Res.* **75**, 171-185.
31. M. van Vliet and A.H.G. Rinnooy Kan (1991). Machine allocation algorithms for job shop manufacturing, *J. Intell. Manufacturing* **2**, 83-94.
32. J. Walrand (1988). *An Introduction to Queueing Networks.* Prentice-Hall, Englewood Cliffs (NJ).
33. Y.-T. Wang and R.J.T. Morris (1985). Load sharing in distributed systems, *IEEE Trans. Comp.* **C-34**, 204-217.
34. H. Yamashita and R.O. Onvural (1994). Allocation of buffer capacities in queueing networks with arbitrary topologies, *Annals of Operations Research* **48**, 313-332.
35. U. Yechiali (1991). Optimal dynamic control of queueing systems, in: *Queueing, Performance and Control in ATM*, eds. J.W. Cohen and C.D. Pack (North-Holland, Amsterdam) pp. 205-217.