The Passport Control Problem or How to Keep a Dynamic Service System Load Balanced?

Alon Itai[*] Michael Rodeh[†] Hadas Shachnai[‡]

Department of Computer Science
The Technion, Haifa 32000, Israel

Abstract

In many real life situations (such as department stores, or passport control booths in airports) parallel queues are formed in front of control stations. Typically, some of the stations are manned while others are not. Classical queuing theory considers the configuration constant, and concentrates on the arrival process. This work explores a new line of research—the case in which the configuration is *dynamic*, and the customers can plan to cope with anticipated changes. Specifically, as the queues build up, management assigns additional officers to the unmanned stations. When this happens—some people move to the newly manned queues from nearby busy queues. In anticipation, people may prefer to line up in busy queues next to unmanned ones.

Mathematically we discuss the problem of dynamic arrangement of the queues in a service system where at any time each server can be in either an active or an inactive mode. A *balancing strategy* determines how customers will be reallocated when a station becomes active. Given a balancing strategy, we seek a partition of customers to queues that minimizes the maximum wait time of a customer in each of the active stations, thereby keeping the system *balanced* at all times. We study two balancing strategies that we call *Split* and *Trim*. For the Split strategy we discuss a special case (the stations are ordered on a line and a single unmanned station is at one end). We show how an optimal partition can be calculated recursively. We then give partitions that approximate the minimal expected wait time within a factor of $1 + O(1/N)$, under each of these strategies, where N is the number of stations. We obtain similar bounds (to within factor 2) for the case, where the number of active servers can be any $1 \leq n \leq N - 1$, and the balancing strategy is Trim.

[*] email: itai@cs.technion.ac.il

[†] Current address: IBM Research Lab. in Haifa, Matam – Advanced Technology Center, Haifa 31905, Israel; email: rodeh@il.ibm.com

[‡] email: hadas@cs.technion.ac.il

Keywords: Queuing, load balancing, response time, service systems.

1 Introduction

Real life queuing systems appear in two major variants:

1. A single line which feeds multiple service stations, in front of which no wait queue exists. This is the common practice in many banks, and in computer systems having several processors, where all user jobs wait in a single ready queue; the scheduler allocates the next available processor to the first job in the queue.

2. Multiple queues, where each queue is served by a single service station. Such systems are used, e.g., in airports, highway toll booths, department stores, and computer systems where the users share several printers, and on sending a job to be printed it is committed to a printer.

We focus on queuing systems of the second type. Classical queuing theory considers the configuration constant, and concentrates on the arrival process. Here we explore unchartered territory—the case in which the configuration is *dynamic*, and the customers can plan to cope with anticipated changes.

Specifically, in the above service systems, as the queues build up, management assigns additional officers to the unmanned stations. When this happens—some people move to the newly manned queue from nearby busy queues. In anticipation, people may prefer to line up in busy queues next to unmanned ones. The problem of interest to us is the expected wait times in the different queues, as a function of the configuration of the manned/unmanned queues and the anticipation of changes in the system.

1.1 The Queuing Model and Problem Statement

Consider a queuing system which consists of N service stations $\{0, \ldots, N-1\}$. Each service station i has a single server, and a wait-queue Q_i, from which customers are picked up in a first-come-first-served order; all stations provide the same type of service, with equal speed, thus an arriving customer can join any of the N queues.

We assume that time is discrete; at any time unit each of the servers can be in either an *active* or an *inactive* mode. An active server i, $0 \leq i \leq N-1$, gives service to the first customer in queue i (whenever the queue is non-empty). The service time of a customer is a single time unit.

We consider a *controlled* queuing system, in which the total number of waiting customers is some constant $L > N$. Suppose there are n active servers at the beginning of some time unit t; by the end of this time unit n customers are serviced and leave the system; then, in the subsequent time unit, n new customers are admitted into the system. Each active queue receives a new arriving customer, so that the queue lengths are stationary, i.e., do not change, as long as the number of servers is *fixed*. Such a situation can be maintained by an external regulator, who is aware of the number of active servers, and regulates the number of new customers. (See Section 1.2.)

Hence we assume a deterministic queuing model, where in each time unit n customers leave the system and n new customers arrive. However, the system is dynamic in the sense that active servers may become inactive, and inactive servers can become active: at time t, each server can change mode with some probability, that depends on the number of active servers in t. When the number of active servers is *fixed*, an arriving customer joins one of the queues, and waits in this queue for service. However, when the number of active servers changes, the customers adjust to the updated system configuration, by moving from one queue to another. In particular, when server i becomes inactive the customers lined up in its queue move to any of the other service stations that are currently active; when an inactive server i becomes active, customers from other active queues move to station i and form a queue there. Thus, at any time t, the number of queues in the system is n, where $1 \leq n \leq N$.

Formally, the state of the servers is represented by the *activity* vector $\mathcal{A} = (a_0, \ldots, a_{N-1})$, where $a_i \in \{0, 1\}$, and $a_i = 0$ (1), if the ith server is inactive (active).

The state of the system at time t is completely determined by the *configuration* $C^t = (\mathcal{A}^t, q^t, pos^t)$: $\mathcal{A}^t$ is an activity vector; for customer c, $q^t(c)$ is the queue that c belongs to, and $pos^t(c)$ the position of c in $q^t(c)$. Thus, if c is a customer about to be served at time t then $pos^t(c) = 1$. The functions q and pos are defined only for customers who are in the system at time t. A configuration C^t induces a *partition* ℓ^t of the L customers: ${\ell_i}^t = |\{c \mid q^t(c) = i\}|$.

A configuration should satisfy the following rules:

1. If $a_i^t = 0$ then $\ell_i^t = 0$.
2. For all i for which $\ell_i^t > 0$, pos^t is a bijection from $\{c \mid q^t(c) = i\}$ to $\{1, \ldots, \ell_i^t\}$.

In the normal course of events no server joins or leaves the system, thus $\mathcal{A}^{t+1} = \mathcal{A}^t$. Each active server serves the first customer, i.e., if $pos^t(c) = 1$ then $pos^{t+1}(c)$ is undefined, and the queues follow a strict FIFO order, that is, if $pos^t(c) > 1$ then $pos^{t+1}(c) = pos^t(c) - 1$. Moreover, $\sum_i a_i^t$ new customers join the system; they join the end of the current queues so as to satisfy rules (1) and (2) above.

The transition of customers (in response to changes in the number of active servers) is determined by a *balancing strategy* $\mathcal{S}$. This strategy defines how to update q and pos. We shall consider only strategies for which the way c is updated depends only on $q(c)$ and $pos(c)$. Since the queues and the strategies follow the FIFO discipline, the effect of the strategy in response to a change in the system is completely determined by the partition.

We study strategies that minimize the movement of customers, and prefer moving customers to adjacent queues, i.e., from Q_i to Q_{i-1} or Q_{i+1}. This models physically separated queues in which movement is restricted to adjacent queues, as well as multi-hop networks, where the cost of moving customers is related to the distance between the servers.

Let $P^+ = \{{P^+}_{jn}\}$, $P^- = \{{P^-}_{jn}\}$ be the $N \times N$ activation/de-activation probability matrices: Given that server j is inactive (active) at time t, ${P^+}_{jn}$ (${P^-}_{jn}$) is the probability that this server becomes active (inactive) in the next time unit, when the number of active servers at time t is n.

Our performance measure is the maximal expected wait time of a customer, which finds the system with activity vector $\mathcal{A}$ and partition ℓ; the maximum is taken over all the active queues. Thus, for any strategy $\mathcal{S}$, there exists a set of (initial) partitions, with which $\mathcal{S}$ performs best. However, finding this set of partitions is computationally hard (see in Section 2.1), therefore we focus on finding a set of partitions that provides close approximation to the optimal under the strategy $\mathcal{S}$.

For a given strategy $\mathcal{S}$ we denote by $\text{max\_W}(\mathcal{S}, \mathcal{A}, \ell, P^+, P^-)$ the maximum expected wait time of a customer arriving when the activity vector is $\mathcal{A}$, and the partition is ℓ. (The maximum is taken over all arriving customers.) We shorten $\text{max\_W}(\mathcal{S}, \mathcal{A}, \ell, P^+, P^-)$ to $\text{max\_W}(\mathcal{S}, \ell)$ when $\mathcal{A}$, P^+ and P^- are understood from the context; $\text{min\_W}(\mathcal{S}, \ell)$ is defined similarly.

Our optimization problem can be stated as follows:

> Given a set of N servers, the matrices P^+, P^-, a strategy $\mathcal{S}$, and an activity vector $\mathcal{A}$, find a partition ℓ such that $\text{max\_W}(\mathcal{S}, \mathcal{A}, \ell, P^+, P^-)$ is minimized.

Note that as long as no server is idle, the average waiting time in the system is equal to the average queue length, independent of the strategy. However, our goal is to minimize the variance of the wait time between customers that join *different* queues. Denote by $W_{OPT}(\mathcal{S}, \mathcal{A})$ the maximal expected wait time of a customer that finds the system with activity vector $\mathcal{A}$, and an *optimal* partition, where the balancing strategy is $\mathcal{S}$. We call $W_{OPT}(\mathcal{S}, \mathcal{A})$ the *min-max* expected wait time.

Definition 1.1 *Given a strategy $\mathcal{S}$ and activity vector $\mathcal{A}$, the partition ℓ yields a $(1+\delta)$-approximation to the min-max expected wait time, if*

$$max\_W(\mathcal{S}, \ell) \leq (1+\delta) W_{OPT}(\mathcal{S}, \mathcal{A}) \ .$$

Our secondary goal is to find a partition that minimizes the balance ratio.

Definition 1.2 *The balance-ratio of a partition ℓ under a balancing strategy $\mathcal{S}$ is given by*

$$\beta(\mathcal{S}, \ell) = \frac{max\_W(\mathcal{S}, \ell)}{min\_W(\mathcal{S}, \ell),} \ .$$

Throughout the paper we refer to a simplified model, in which the probability that any active server becomes inactive is negligible, thus the expected wait time of any customer depends only on activation of servers while this customer is in the system, and the resulting change in the position of the customer in the wait queue. The analysis of the more general case, where transitions of customers can occur due to de-activation of servers, is left for future work.

1.2 Application to Communication Systems

Consider a communication system consisting of N parallel lines between s and t, of which only some are active. Due to external considerations (such as cost or response time) the router wishes to send as many messages (customers) as possible through the system. However, since the router has a buffer which can hold at most L messages, it should send a new message, only when an old one has left the system. If n sub-lines are active, at each time unit, n messages leave the line, hence the router can release only n new messages in every time unit.

When a non-active line becomes active then the router can move to it messages from other lines. Thus new messages will be sent only after the messages currently waiting were sent. The router can also rearrange all the messages so as to preserve the FIFO discipline. Note, however, that if moving message among queues causes delays, the router might want to minimize such moves, as well as distribute the messaged between the lines in anticipation of the repair.

1.3 Related Work

In the context of queuing theory, a model close to the one discussed above is the *server of the walking type* model, where each server in the system can become inactive for a certain amount of time, upon completion of a service period [4, 3, 7, 8]. Related work on this model studied the steady state distribution on queue lengths and on the wait times of customers, when some probabilistic assumptions are used, e.g., for determining the length of the "vacation" taken by a server. However, no balancing strategies are used in this model, i.e., when the server is inactive the customers are assumed to line up in its service station, until the server becomes active again.

Another related problem is the dynamic server problem introduced in [1]. This problem refers to a task system in which the number of available servers can change at any time. The paper presents competitive algorithms for this problem; the model is slightly different than our service system, as customers do not line up in queues for service; they are distributed in many sites in a computer network, thus rearrangements of the queues are inapplicable. A similar dynamic server model was used also in the study of load balancing schemes for multimedia systems (see in [5, 6]).

1.4 Main Results

We consider two natural balancing strategies, that attempt to minimize the movement of customers: the *Split* (see Section 2), and the *Trim* strategy (see Sections 3 and 4). We derive efficient initial partitions for these strategies. In particular,

- For the Split strategy, if $n = N - 1$ then
  - (i) We show how an optimal partition can be calculated recursively, based on the knowledge of the activation probability of the inactive server.
  - (ii) We give a partition which yields a $1 + O(1/N)$-approximation to the min-max expected wait time.
- For the Trim strategy:
  - (i) For $N/2 \leq n \leq N - 1$ we give a set of partitions that yield a $(1 + O(1/N))$-approximation to the min-max expected wait time.
  - (ii) For $0 < n < N/2$ we give a set of partitions that yield a $(2 + \varepsilon)$-approximation to the min-max expected wait time, where $\varepsilon = o(1)$.

1.5 Organization of the Paper

In Sections 2 and 3 we derive results for the case where $n = N-1$. In Section 2 we present the Split strategy, for which we show (in Section 2.1) how an optimal partition can be computed recursively. We then give (in Section 2.2) a partition that is a close approximation to the optimal. In Section 3 we propose the Trim strategy, and give an efficient partition for the case $n = N - 1$.

In Section 4 we show how our results can be extended to apply to the case, where the number of inactive servers in the system is any $0 < n \leq N - 1$. Finally, we summarize in Section 5, with some directions for future work.

2 The Split Strategy

Assume first that the servers are linearly ordered, Q_0 is inactive, and ℓ_i customers wait at Q_i, $i = 1, \ldots, N-1$. The probability that a server appears at Q_0 in the next time unit is ${P^+}_{0(N-1)}$. Since system configuration can change only due to the activation of a server in Q_0, we note that the arrival probability of this server reflects a geometrically distributed *repair time*, with probability $p = {P^+}_{0(N-1)}$. Let $\bar{L} = L/N$ be the average queue length when all the stations are active.

In this section we discuss a balancing strategy that we call *Split*. We first show how an optimal partition can be calculated recursively. We then show that if $p > 1/\bar{L}$ then a simple partition (based on accumulating a large amount of customers near the inactive server) can yield expected wait times that are close to the optimal within a factor of $1 + O(1/(Lp)) = 1 + O(1/N)$.

The Split balancing strategy is as follows:

At step $i = 1$ (when a server arrives at Q_0), every other customer from Q_1 moves to Q_0. After that, Q_0 has $\bar{L}_0 = \lfloor \ell_1/2 \rfloor$ customers and Q_1 has $\bar{L}_1 = \lceil \ell_1/2 \rceil$ customers.

At step $i > 1$: queues $Q_0, \ldots, Q_{i-1}$ have each $\bar{L}_{i-1} = \frac{1}{i}\sum_{j=1}^{i-1} \ell_j$ customers. The customers of Q_i who have to wait more than $\bar{L}_{i-1}$ time, evenly split to queues $Q_0, \ldots, Q_i$ in a cyclic manner, i.e., customer $\bar{L}_{i-1} + 1$ moves to Q_0,..., customer $\bar{L}_{i-1} + i$ moves to Q_{i-1}, customer $\bar{L}_{i-1} + i + 1$ remains in Q_i, customer $\bar{L}_{i-1} + i + 2$ moves to Q_0 etc.

Thus, if $q(c) = i$ then in the new configuration $q'(c)$ and $pos'(c)$ are defined as follows:

$$q'(c) = \begin{cases} i & pos(c) \leq \bar{L}_{i-1} \\ (pos(c) - \bar{L}_{i-1} - 1) \bmod (i+1) & \text{otherwise.} \end{cases}$$

and

$$pos'(c) = \left|\{d \mid q'(d) = q'(c) \text{ and } (q(d), pos(d)) <_{\text{lex}} (q(c), pos(c))\}\right|,$$

where $(a, b) <_{\text{lex}} (a', b')$ if $a < a'$ or $a = a'$ and $b < b'$. Note that in the new configuration all queues have length $\ell'_i \in \left\{\left\lfloor \bar{L} \right\rfloor, \left\lceil \bar{L} \right\rceil\right\}$.

2.1 Finding an Optimal Partition

Assume that the arrival probability of a server at Q_0 is a fixed parameter p. For any $t \geq 1$, the probability that the server arrives at Q_0 at time t is

$$p_t = (1-p)^{t-1} p \quad . \tag{1}$$

Let $W_i(Split, \ell)$ be the expected wait time of a customer at Q_i, given that the initial lengths of the queues are $\ell = (0, \ell_1, \ldots, \ell_{N-1})$.

Let $\Delta_i = \ell_i - \bar{L}_{i-1} = \ell_i - \frac{1}{i}\sum_{j<i} \ell_j$ denote the excess number of customers in Q_i over the average number of customers in queues $Q_0, \ldots, Q_{i-1}$, after the customers of these queues moved (in response to the arrival of the new server at Q_0), but just before the customers of Q_i moved. Consider a customer c that arrives at time $t = 0$ and joins Q_i. We simplify the calculations by allowing the queue lengths to be non-integral numbers.

If the server arrives at time t, $1 \leq t \leq \Delta_i$, then $\frac{i}{i+1}(\Delta_i - (t-1))$ customers of Q_i that are before c leave Q_i to join $Q_0, \ldots, Q_{i-1}$. Thus the wait time is reduced by that amount.

We will use the equality

$$\sum_{t=1}^{k} p_t t = p \sum_{t=1}^{k} (1-p)^{t-1} t = \frac{1}{p}\left[1-(1-p)^k(kp+1)\right] \ . \tag{2}$$

The expected wait time at Q_i is

$$\begin{aligned} W_i(\ell, Split) &= \sum_{t=1}^{\Delta_i} p_t \left(\ell_i - (\Delta_i - t + 1)\frac{i}{i+1}\right) + \sum_{t>\Delta_i} p_t \ell_i \\ &= \ell_i - \frac{i}{i+1}\left[\Delta_i \sum_{t=1}^{\Delta_i} p_t - \sum_{t=1}^{\Delta_i} t p_t + \sum_{t>\Delta_i} p_t \ell_i\right] \\ &= \ell_i - \frac{i}{i+1}\left[\Delta_i \left(1-(1-p)^{\Delta_i}\right) - \frac{1}{p}\left(1-(1-p)^{\Delta_i}(\Delta_i p + 1)\right) + 1 - (1-p)^{\Delta_i}\right] \\ &= \ell_i - \frac{i}{i+1}\left[\Delta_i - \frac{1}{p} + \frac{1}{p}(1-p)^{\Delta_i} + 1 - (1-p)^{\Delta_i}\right] \ . \end{aligned}$$

Given queues of length $\ell_1, \dots, \ell_{N-1}$, a new customer joins the queue with minimum expected waiting time. Hence the system will become totally balanced when

$$W_1(Split, \ell) = W_2(Split, \ell) = \cdots = W_{N-1}(Split, \ell) \ .$$

Let ℓ_1 be given, we first calculate $W_1(Split, \ell)$ and choose ℓ_2 such that $W_1(Split, \ell) = W_2(Split, \ell)$. We can easily perform these calculations, since $W_2(Split, \ell)$ depends on p and on $\Delta_2 = \ell_2 - \ell_1/2$. We continue in this fashion to compute the values of $\ell_3, \dots, \ell_{N-1}$. The key observation is that ℓ_i depends only on p and $\Delta_i = \ell_i - \frac{1}{i}\sum_{j<i} \ell_j$. Hence ℓ_i depends only on p and $\ell_1, \dots, \ell_{i-1}$, which we have already computed.

If only L, the total number of customers, is known we can perform a binary search on ℓ_1 to find values $\ell_1, \dots, \ell_{N-1}$ that balance the system and satisfy $\sum_{i=1}^{N-1} \ell_i = L$.

Note that for $p = 1$, we get

$$W_i(Split, \ell)_{|p=1} = \ell_i - \frac{i}{i+1}\Delta_i = \ell_i - \frac{i}{i+1}(\ell_i - \frac{1}{i}\sum_{j<i} \ell_j) = \frac{1}{i+1}\sum_{j=1}^{i} \ell_j = \bar{L}_i \ .$$

Since $\bar{L}_1 = \ell_1/2$, in this case the system is balanced when $\ell_1/2 = \ell_2 = \cdots = \ell_{N-1}$. In other words, the first queue (next to the inactive server) has twice as many customers in it—since with probability $p = 1$ at time $t = 1$ a server appears at Q_0, half of the waiting customers move to Q_0; then, all queues are of equal length.

2.2 Approximating the Min-max Expected Wait Time

Assume now that the inactive server is at position h, for some $0 \leq h < N-1$. Thus, the activity vector is $\mathcal{A} = (1, 1, \ldots, 0, \ldots, 1)$, with $a_i = 0$ for $i = h$, and $a_i = 1$ otherwise. In the following we show how the min-max expected wait time can be closely approximated for this case.

We propose to use a static partition $(\ell_0, \ldots, \ell_h, \ldots, \ell_{N-1})$, in which all the queues that are of the same distance from the inactive station have the same length. In fact, we show that a close approximation to the optimum can be obtained by choosing a partition, in which the queues next to the inactive server have the same length ℓ', and all the other queues in the system have the same length ℓ''. Hence, we only need to compute the ratio $\ell'/\ell'' \geq 1$.

When the inactive server is positioned at the end of the line, the queue will be balanced (upon activation of this server) by the Split strategy. When the server is at position h, for some $0 < h < N-1$, the queues will be balanced using the Split strategy with a slight modification: the balancing process will handle in phase i all queues that are of distance at most i from the inactive server. Thus, after the first phase queues $h-1, h, h+1$ will be of equal length. After the second phase queues $h-2, h-1, h, h+1, h+2$ will be of equal length, and so on.

Let $s \in \{1, 2\}$ denote the number of immediate neighbors of the inactive server. For $p \geq 1/\bar{L}$ we define the partition $\ell^{Split} = (\ell_0, \ldots, \ell_{N-1})$, where

$$\ell_i = \begin{cases} \frac{1}{s}(s+1)\bar{L} + \frac{1}{p}\left(\frac{1}{p(N-1)} - \frac{1}{s}\right) & i \text{ is a neighbor of inactive server} \\ 0 & i = h \\ \bar{L} + 1/(p(N-1)) & \text{otherwise.} \end{cases}$$

Theorem 2.1 *The partition ℓ^{Split} yields a $(1 + 1/(\bar{L}\ p(N-1)))$- approximation to the min-max expected wait time, when the balancing strategy is Split, that is,*

$$W_i(\text{Split}, \ell^{Split}) \leq \left(1 + \frac{1}{\bar{L}\ p(N-1)}\right) W_{OPT}(\text{Split}, \mathcal{A}) \qquad \forall\ 0 \leq i \leq N-1\ . \tag{3}$$

Proof: Let p_t be as defined in (1).

(i) The case $h = 0$. Suppose a server arrives to queue h at time $t \leq \ell_1$. At that time a customer that arrived at time 0 is at position $\ell_1 - t$, half of the customers before it move to queue h, thus its remaining time decreases by $(\ell_1 - t)/2$. If the server arrives at queue h after time ℓ_1, the customer is already served, thus its waiting time is not decreased—it remains ℓ_1. The expected waiting time is therefore,

$$W_1(Split, \ell^{Split}) = \ell_1 - \sum_{t=1}^{\ell_1} p_t \frac{\ell_1 - t}{2}$$

$$
\begin{aligned}
&= \ell_1 - \frac{1}{2}\left(1-(1-p)^{\ell_1}\right)\ell_1 + \frac{1}{2}\sum_{t=1}^{\ell_1} p_t t \\
&= \frac{1}{2}\ell_1 + \frac{1}{2}(1-p)^{\ell_1}\ell_1 + \frac{1}{2}\sum_{t=1}^{\ell_1} p_t t \quad .
\end{aligned}
$$

Using equation (2) we have

$$
\begin{aligned}
W_1(Split, \ell^{Split}) &= \frac{1}{2}\ell_1 + \frac{1}{2}(1-p)^{\ell_1}\ell_1 + \frac{1}{2p}\left[1-(1-p)^{\ell_1}(\ell_1 p+1)\right] \\
&= \frac{1}{2}\ell_1 + \frac{1}{2}(1-p)^{\ell_1}\ell_1 + \frac{1}{2p}\left[1-(1-p)^{\ell_1}\ell_1 p-(1-p)^{\ell_1}\right] \\
&= \frac{1}{2}\ell_1 + \frac{1}{2p} - \frac{1}{2p}(1-p)^{\ell_1} \; .
\end{aligned}
$$

Taking $\ell_1 = 2\bar{L} + \frac{1}{p}(\frac{1}{N-1} - 1)$, we get

$$
\begin{aligned}
W_1(Split, \ell^{Split}) &\leq \frac{1}{2}(2\bar{L} + \frac{1}{p}(\frac{1}{(N-1)} - 1)) + \frac{1}{2p} \\
&= \bar{L}(1 + \frac{1}{2\bar{L}p(N-1)})
\end{aligned}
$$

which yields inequality (3).

(ii) For the case in which $0 < h < N-1$, we have $W_{h+1}(Split, \ell^{Split}) = W_{h-1}(Split, \ell^{Split})$. Until the arrival of a server at queue h, queues $h-1$ and $h+1$ have the same length $\ell = \ell_{h-1} = \ell_{h+1}$. A customer c that arrived at time 0 to queue $h-1$, will be at time t at position $\ell - t$. Upon the arrival of the new server at queue h, third of the customers before c in Q_{h-1} will move to Q_h so that the three queues have the same length. Thus the waiting time of customer c will decrease by $\frac{1}{3}(\ell - t)$.

$$
\begin{aligned}
W_{h-1}(Split, \ell^{Split}) &= \ell - \sum_{t=1}^{\ell} p_t \frac{1}{3}(\ell - t) \\
&= \ell - \frac{\ell}{3}\sum_{t=1}^{\ell} p_t + \frac{1}{3}\sum_{t=1}^{\ell} t p_t \\
&= \ell - \frac{\ell}{3}\left(1-(1-p)^{\ell}\right) + \frac{1}{3p}\left[1-(1-p)^{\ell}(\ell p+1)\right] \\
&= \frac{2}{3}\ell + \frac{1}{3}\ell(1-p)^{\ell} + \frac{1}{3p}\left[1-(1-p)^{\ell}\ell p-(1-p)^{\ell}\right] \\
&= \frac{2}{3}\ell + \frac{1}{3p}\left[1-(1-p)^{\ell}\right] \\
&= \frac{1}{3}\left(2\ell + \frac{1}{p}\right) - \frac{(1-p)^{\ell}}{3p}
\end{aligned}
$$

Substituting in the last equation ℓ by $\frac{1}{2}(2\bar{L}+\frac{1}{p}(\frac{1}{N-1}-\frac{1}{2})$, we have the statement of the theorem.

□

3 The Trim Strategy

In this section we propose a balancing strategy called Trim. We define a partition, that is shown to provide a close approximation to the min-max expected wait time, while maximizing the balance in the system.

The following is an informal description of the strategy. Consider first the case, where the inactive server is at the end of the line (i.e., $h = 0$). Suppose that initially there are ℓ_i customers in queue i, $1 \leq i \leq N-1$. When server 0 becomes active, customers from queues 1 through $N-1$ transfer to Q_0 as follows.

> If for a customer c, $q(c) = i$ and $pos(c) = j$, then all customers in queues $Q_1, \ldots, Q_i - 1$ whose positions are larger that $\bar{L}$ will line up in Q_0 before c. In addition, any customer d for which $q(d) = i$ and $pos(d) < j$ will also precede d in Q_0.

We call this balancing strategy *Trim*, since it trims all the long queues in the system, and transfers the excess customers to Q_0, until all queues are of the same length $\bar{L}$.

Formally,

$$q'(c) = \begin{cases} q(c) & pos(c) \leq \bar{L} \\ 0 & \text{otherwise.} \end{cases}$$

and

$$pos'(c) = \left|\{d \mid q'(d) = q'(c) \text{ and } (q(d), pos(d)) <_{\text{lex}} (q(c), pos(c))\}\right|.$$

Note that if $q'(c) = q(c)$ then $pos'(c) = pos(c)$, and as with the Split strategy, in the new configuration all queues have length $\ell'_i \in \left\{\lfloor \bar{L} \rfloor, \lceil \bar{L} \rceil\right\}$.

Assume now, that the inactive server is located at station h, for some $1 \leq h \leq N-2$. For this case we slightly modify the Trim strategy: if a customer c is at distance i from the inactive server, then when the server becomes active

- All the customers at positions $j > \bar{L}$, that belong to queues closer than i will line-up in Q_h before c.
- All the customers at positions $j > \bar{L}$ who belong to the queues $\{h-i, h+i\}$ then join in random order at the end of Q_h, including c.

We call this strategy *Random Trim.*

In the next result we define a static partition of the customers to queues, which is shown to be efficient with respect to our two measures (min-max wait time and balance ratio), when the balancing strategy is (Random) Trim.

Let the random variable X denote the activation time of the server, then $X \sim G(p)$ (i.e., X is geometrically distributed, and $\text{Prob}(X = t) = p(1-p)^{t-1}$).

Theorem 3.1 *Let M be the median of the random variable X, and let*

$$r = \lg_2(\bar{L}\ /M) \quad , \tag{4}$$

and

$$r^* = \max\left\{ r' \geq r \mid r' \leq \frac{2^{r'}-1}{2^{r'-r}} \right\} \quad . \tag{5}$$

Define h' as follows: if $h = 0$ then $h' = 1$, otherwise $h' = h-1$. Denote by ℓ^{Trim} the partition $(\ell_0, \ldots, \ell_{N-1})$, where

$$\ell_i = \begin{cases} 0 & i = h \\ \bar{L}(2 - 1/2^{r^*}) & i = h' \\ \bar{L}(1 + 1/(2^{r^*}(N-2))) & \textit{otherwise.} \end{cases}$$

Then for any $r \geq 1$ and a system of $N > 2$ stations,

(i) ℓ^{Trim} yields a $(1 + \frac{1}{2^{r^*}(N-2)})$-approximation to the min-max expected wait time.

(ii) $$\beta(\mathit{Trim}, \ell^{Trim}) \leq \frac{1 + \left(2^{r^*}(N-2)\right)^{-1}}{1 - 2^{-r^*}} \quad .$$

Proof: We give the proof for $h = 0$. The generalization to arbitrary h is straightforward. We first note, that if $M \leq \bar{L}/2$ there exists $r^* \geq r$ satisfying (5) (since clearly, $r^* = r$ satisfies the inequality). Now, we show separately the approximation-bound and the balance-ratio of the partition ℓ^{Trim}.

(i) We first show, that

$$W_i(\text{Trim}, \ell^{Trim}) \leq (1 + \frac{1}{2^{r^*}(N-2)}) W_{OPT}(\text{Trim}, \mathcal{A}) \qquad \forall\ 1 \leq i \leq N-1 \ . \tag{6}$$

For ${\ell'}_1 = \bar{L} \cdot (1 - 1/2^{r^*})$, let B be the event

“The inactive server becomes active within less than ℓ'_1 time units” , (7) then

$$\begin{aligned} W_1(\text{Trim}, \ell^{Trim}) &= \text{Prob}(B)\ell_1' + (\bar{L} + \ell_1')(1 - \text{Prob}(B)) \\ &= \ell_1' + \bar{L}(1 - \text{Prob}(B)) \ . \end{aligned}$$

Since $M = \bar{L}/2^r$ and r^* satisfies (5), the probability that the server remains inactive in the next ℓ_1' time units is

$$1 - \text{Prob}(B) = (1-p)^{\bar{L}/2^r \cdot \frac{2^{r^*}-1}{2^{r^*-r}}} \leq \frac{1}{2^{r^*}} \ . \tag{8}$$

Hence,

$$W_1(\text{Trim}, \ell^{Trim}) = \bar{L}(1 - \frac{1}{2^{r^*}}) + \bar{L}\frac{1}{2^{r^*}} \leq \bar{L} \ .$$

Since $W_{OPT}(\text{Trim}, \mathcal{A}) \geq \bar{L}$, obviously inequality (6) holds for $i = 1$.
For queues $i = 2, \ldots, N-1$ we have

$$\ell_i = \bar{L}\left(1 + \frac{1}{2^{r*}(N-2)}\right) \ ,$$

and again inequality (6) holds.

(ii) We now bound the balance-ratio of the system under Trim. Since for any $1 \leq i \leq N-1$,

$$W_i(\text{Trim}, \ell^{Trim}) \leq \left(1 + \frac{1}{2^{r^*}(N-2)}\right)\bar{L} \ ,$$

it remains to show that

$$\text{min\_W}(\text{Trim}, \ell^{Trim}) \geq \bar{L}\left(1 - \frac{1}{2^{r^*}}\right) \ .$$

We note, that the choice of ℓ_1 implies

$$W_1(\text{Trim}, \ell^{Trim}) \geq \bar{L}\left(1 - \frac{1}{2^{r^*}}\right) \ .$$

Let C be the event

"The inactive server becomes active within the next $\bar{L}/(2^{r^*}(N-2))$ time units".

Then for $i = 2, \ldots, N-1$

$$W_i(\text{Trim}, \ell^{Trim}) \ \geq \text{Prob}(C)(\ell_1 - \bar{L}) + (1 - \text{Prob}(C))\bar{L} \geq \ell_1' \ ,$$

which yields the statement of the theorem.

□

Corollary 3.2 *If $p \geq \frac{3}{L}$ then the partition ℓ^{Trim} defined in Theorem 3.1 gives*

(i) a $\left(1 + \frac{1}{8(N-2)}\right)$-approximation to the min-max expected wait time.

(ii) $\beta(Trim, \ell^{Trim}) \leq \frac{9}{7}$.

4 Extensions

In this section we discuss the case, where the number of active servers is any $1 \leq n \leq N-1$. We show how our results for the case $n = N-1$ can be used to obtain similar bounds for the Split and the Trim strategies. We exemplify this generalization for the Trim strategy.

a. If $n \geq N/2$, then we can divide the set of queues to $N-n$ *regions*, such that each region $1 \leq j \leq K$ contains exactly one inactive station, and $A_j \geq 1$ active stations, where $\sum_{j=1}^{N-n} A_j = n$ (see Figure 1). We call this version of the Trim strategy the *Region-Trim*. Note, that the stations in each region are not necessarily neighboring stations on the line. The total number of customers in the queues of region j is $A_j \cdot \bar{L}$.

   In region j, $1 \leq j \leq N-n$, the partition of the customers to the queues will be defined as for the Trim strategy, namely, the partition of $L_j = A_j \bar{L}$ customers, in a system of $N_j = A_j + 1$ queues, in which a single server is inactive. Denote by h_j the jth inactive server, and let h'_j be the active station closest to h_j in region j, then the partition of Region-Trim is given by $\ell^{RT} = (\ell_0, \ldots, \ell_{N-1})$, where

$$\ell_i = \begin{cases} 0 & i = h_j \text{ for some } 1 \leq j \leq N-n \\ \bar{L}(2 - 1/2^{r^*}) & i = h'_j \text{ for some } 1 \leq j \leq N-n \\ \bar{L}(1 + 1/(2^{r^*}(N-2))) & \text{otherwise,} \end{cases}$$

   where r^* is defined in (5).

b. For the case where $K > N/2$ we choose i^*—one of the active queues in the system $(0 \leq i^* \leq N-1)$—to be the *master queue*; Q_{i^*} will consist of $\bar{L}\ (N-(2-1/2^{r^*})(n-1))$ customers. Upon activation of a server at Q_{h_j}, for some $1 \leq j \leq N-n$, the last $\bar{L}(2-1/2^{r^*})$ customers in Q_{i^*} move to Q_{h_j} (see Figure 2). Note, that when n reaches $N/2$, Master-Trim becomes Region-Trim.

   Thus, the partition for Master-Trim is given by $\ell^{MT} = (\ell_0, \ldots, \ell_{N-1})$, where

$$\ell_i = \begin{cases} 0 & i = h_j \text{ for some } 1 \leq j \leq N-n. \\ \bar{L}\ (N - (2 - 1/2^{r^*})(n-1)) & i = i^* \\ \bar{L}(2 - 1/2^{r^*}) & \text{otherwise.} \end{cases}$$

   where r^* is defined in (5).

   Assume that the probability that the jth inactive server becomes active in the next time unit is a monotonically increasing function of the number of inactive servers, and that ${P^+}_{j(N-1)} \geq 1/\bar{L}$. In the following we show, that for any $n \in [1, N-1]$, the expected wait time of these partitions under the Trim strategies is at most $2+\varepsilon$ times the optimal, where $0 < \varepsilon < 1$ is small.

   Theorem 4.1 *Let M and r be defined as in (4). For a system of $N > 2$ stations, and any $r \geq 1$,*

a. *If $N/2 \leq n \leq N-1$, then ℓ^{RT} yields a $\left(1+\frac{1}{2^{r^*}(N-2)}\right)$-approximation to the min-max expected wait time.*

b. *If $1 \leq n < N/2$, then ℓ^{MT} yields a $(2+\varepsilon)$-approximation to the min-max expected wait time, where $\varepsilon = o(1)$.*

Proof: Note, that we compare the expected wait time under the two variants of Trim to the expected wait time under an *optimal* partition, which uses *any* Trim-based strategy, for balancing the system.

a. For the case where $N/2 \leq n \leq N-1$, we can obtain the bound for each region separately, since the servers are independent. Hence, the statement of the theorem follows from Theorem 3.1.

b. For the case where $1 \leq n \leq N/2$, each customer joins either the master queue, or one of the active stations. We need to show, that for any $0 \leq i \leq N-1$
$$W_i(\text{Master-Trim}, \ell^{MT}) \leq (2+\varepsilon) W_{OPT}(\text{Trim}, \mathcal{A}), \tag{9}$$
where $\varepsilon = o(1)$.
We now show, that (9) holds for $i = i^*$. As before, h_j is the jth inactive server, $1 \leq j \leq N-n$. Let $Y_1, \ldots, Y_{(N-n)}$ be a sequence of random variables, where
$$Y_j = \begin{cases} 1 & \text{if } h_j \text{ becomes active by } t = \bar{L}(2-1/2^{r^*}) \\ 0 & \text{otherwise,} \end{cases}$$
and let $Y = \sum_{j=1}^{K} Y_j$. Recall that $P^+{}_{jn} \geq 1/\bar{L}$, for any $1 \leq n \leq N-1$. Then
$$\begin{aligned} E[Y] &= \sum_{j=1}^{N-n} \text{Prob}(Y_j = 1) \\ &\geq (N-n)(1-e^{-2}) \equiv \mu > \frac{3}{4}(N-n) \ . \end{aligned}$$
By Chernoff's inequality [2], for any $0 < \delta < 1$
$$\text{Prob}(Y < (1-\delta)\mu) < e^{\frac{-\mu\delta^2}{2}} \ .$$
Taking $\delta = 1/3$, we have
$$\text{Prob}\left(Y < \frac{N-n}{2}\right) < e^{-\frac{N-n}{24}} \equiv \varepsilon_1$$
Therefore, with probability $1-\varepsilon_1$, within $2\bar{L}$ time units half of the inactive servers become active. Thus,
$$\begin{aligned} W_i(\text{Master-Trim}, \ell^{MT}) &\leq 2\bar{L}\ (1-\varepsilon_1) + \ell_{i^*}\varepsilon_1 \\ &\leq 2\bar{L}\ + \varepsilon \end{aligned}$$
where $\varepsilon = o(1)$. Clearly, for any active server, $i \neq i^*$, inequality (9) is satisfied.

□

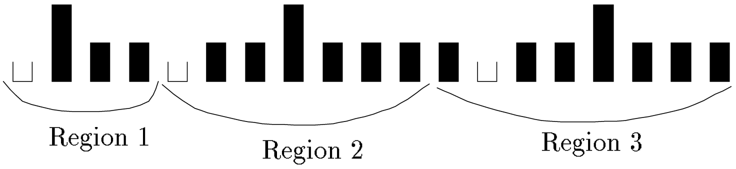

Figure 1: The Region-Trim with $n = N - 3$.

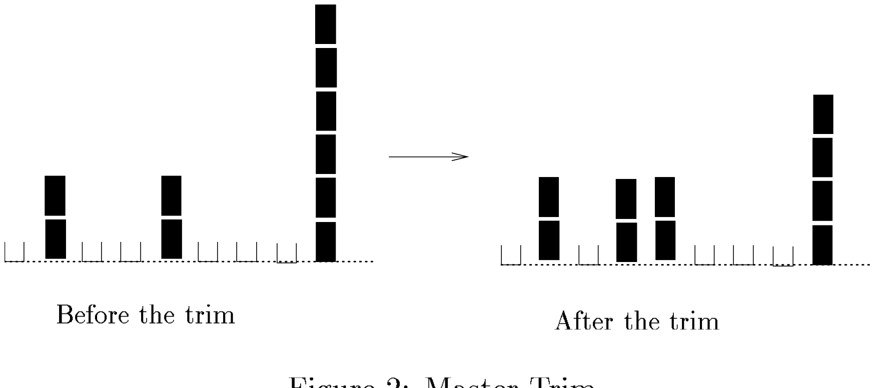

Figure 2: Master-Trim

5 Discussion

In this paper we have studied balancing strategies for a *dynamic* service system model, which has many applications in everyday service systems, as well as in computer and communication systems. We considered two natural balancing strategies for such systems, and gave static partitions, which minimize the maximal expected wait times of customers under these strategies.

Several interesting questions remain open: for simplicity, we assumed a deterministic queuing model. A natural extension is to consider more general distributions on the arrival of customers and the service times. (Note, however, that also in a nondeterministic setting, the arrival rate should be adjusted to the number of active servers, so as not to exceed the service rate). We have studied the *line* topology, which is typical, e.g., in drugstores. It is natural to examine more general topologies, that would represent the application our model, e.g., to process migration in distributed system, in which some nodes (processors) may be inoperational at certain times. Finally, de-activation of servers should be incorporated in the model, i.e., it is reasonable to assume, that any active server can become inactive with some probability, which depends on the id of this server, as well as the system configuration.

References

[1] Charikar M., Halperin D. and Motwani R., "The Dynamic Servers Problem", *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'98)*, San Francisco, January 1998.

[2] Chernoff H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *Annals of Mathematical Statistics*, 23, pp. 493–509, 1952.

[3] Gelenbe, E., Mitrani I., Analysis and Synthesis of Computer Systems, Academic Press, 1980 (Ch. 2, Section 2.2).

[4] Gelenbe E. and Iasnogorodski R., "A Queue with Server of Walking Type", *in Annales de l'Institut Henry Poincaré, Série B (Probabilités et Statistiques)*, Vol. XVI, no. 1, 63-73, 1980.

[5] Golubchik L. and Lui J. C. S., "Bounding of Performance Measures for a Threshold-based Queuing System with Hysteresis", *Proceedings of the ACM SIGMETRICS Conference*, pp. 147–157, 1997.

[6] Lie P. W. K., Lui J. C. S. and Golubchik L., "Threshold-Based Dynamic Replication in Large-Scale Video-on-Demand Systems", *Proceedings of the Eighth International*

Workshop on Research Issues in Database Engineering (RIDE), Orlando, February 1998.

[7] Shachnai, H., Yu P. S., “On Analytic Modeling of Multimedia Batching Schemes”, *Proceedings of the 3rd International Workshop on Multimedia Information Systems (MIS’97)*, Como, September 1997.

[8] Skinner C.E., “A priority queuing model with server of walking type”, *Operations Research*, Vol. 15, pp 278-285, 1967.