Admission Control in Stochastic Event Graphs

Eitan ALTMAN* Bruno GAUJAL† Arie HORDIJK‡

April 27, 1999

Abstract

In this paper, we first show that the expectation of any convex increasing function of the workload (or waiting time) in (max,+) linear systems, under a single input sequence, is multimodular. This is done using a coupling argument and a vectorial version of Lindley's equation. In the second part of the paper, we use this result and the optimization theory based on multimodular costs, to construct the optimal deterministic admission control in general (max,+) linear systems under admission rate constraints. This optimization result only requires stationarity assumptions on the arrival process and on the service times of the servers in the system.

Keywords Multimodularity, (max,+) algebra, event graphs, optimal control.

1 Introduction

The work on admission control in queuing systems can be split into two main domains, depending on the the information available for control decisions. When some knowledge on the state is available (closed loop systems), then the optimal control policy is usually based on dynamic programming techniques [9, 13] and more recently in [14]. When no state information is available (open loop control), then the control is usually static and decisions are taken according to general properties of the cost functions [15]. This paper falls in between since it belongs to the open loop framework but remains dynamic in some sense. Indeed, control decisions will be taken at each arrival epoch.

The general result which is proved in this paper is that when the admission rate is fixed, then the individual arrivals have to be distributed evenly over time in order to minimize the average waiting time (or the workload). The property of the cost function used here in order to prove the optimization result is *multimodularity*. In his seminal paper [10], B. Hajek introduced a convexity property, which he called multimodularity. A multimodular function is only defined in the integer points of a multi-dimensional state space (see Section 3.1 for its definition). However, a multimodular function f (defined on $\mathbb{Z}^m$) can be extended to a function $\tilde{f}$ on $\mathbb{R}^m$. This is done by paving the space $\mathbb{R}^m$ into simplexes which have their extreme points in $\mathbb{Z}^m$. The extension $\tilde{f}$ coincides with f on the integer point and is the linear interpolation of f on each simplex. In [2], it is shown that $\tilde{f}$ is convex if and only if f is multimodular. In the second part of his paper, multimodularity is used by Hajek to analyse how to split an arrival stream of jobs to an exponential server, such that the average expected number of jobs waiting in the queue is minimized. The class of admission policies considered have to accept at least an asymptotic fraction p of the arrivals.

*INRIA, BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: altman@sophia.inria.fr. URL:http://www.inria.fr:80/mistral/personnel/Eitan.Altman/me.html

†INRIA/CNRS/UNSA, BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: gaujal@sophia.inria.fr.

‡Dept. of Mathematics and Computer Science, Leiden University, P.O.Box 9512, 2300RA Leiden, The Netherlands. E-mail: hordijk@wi.leidenuniv.nl. The research of Arie Hordijk was done while he was on sabbatical leave at INRIA, Sophia-Antipolis; it has been partially supported by the Ministère Français de l'Éducation Nationale et de l'Enseignement Supérieur et de la Recherche.

The key property for the analysis in [10] is multimodularity of the number of jobs in the queue at arrival epochs that Hajek used to derive the optimality of the so-called balanced policy. The main ingredient in the proofs are Jensen's inequality applied to the extension of a multimodular function, and Weyl's ergodic theorem.

In this paper we have extended Hajek's results in four different directions.

1. Instead of having one server, we consider a queuing network with one input node. This network is assumed to be a stochastic event graph (queues in tandem for example, fit in this framework, see Section 2).
2. Instead of having exponentially distributed service times, the service times of the customers admitted in the network may be any stationary process.
3. Besides the average workload up to the various nodes in the network, we also consider the average traveling time (*i.e.* the average waiting time in case of one server).
4. More generally, any performance measure which is a weighted average of expectations of convex functions of the workload (traveling time) to any of the queues can be considered (note that each convex function depends on the measure for one of the queues and no multi-dimensional convex function is considered, see Lemmas 7.4 and 7.6 for a precise formulation).

For this more general situation, we show that the same result holds, *i.e.* under all admission policies, with an asymptotic fraction p of acceptance, the balanced policy with rate p is optimal. This policy can be given by:

$$a_n = \lfloor np \rfloor - \lfloor (n-1)p \rfloor,$$

where $\lfloor X \rfloor$ denotes the largest integer smaller or equal to X, and $a_n = 1$ (0) if the n-th arrival is accepted (rejected). The intuitive idea behind it is that for stationary arrivals, balanced admissions spread out the entrance times of the admitted jobs "most equally".

Our analysis has two parts. In the first part (done in reference [2]), we establish a rather technical asymptotic analysis which shows that the conditions of Lemma 7.4 and Lemma 7.6 are sufficient for the optimality of the balanced policy in case of the average workload (traveling time).

What remains then, is to verify the conditions of Lemma 7.4 (Lemma 7.6). The main point is the verification of the multimodularity property, its proof for the single server case relies on the Lindley equation. For the stochastic event graph, it is possible to derive a generalized version of this equation (see [6] and Section 5.2). Although technically and notationally more involved, it turns out that the verification of multimodularity in this far more general setting has the same structure as for the one-server queue with a general service distribution and "first in first out" service discipline. Via a counter example, we point out in Section 3.2 that the multimodularity property does not hold sample path wise (when we consider a deterministic sequence of service times which are not equal). Therefore, we consider the expected workloads (expected traveling times) and we apply a coupling of the service times (inter-arrival times). See Section 4 for details.

The paper is structured as follows. In Section 2, we present an overview of the model that will be studied, that is (max,+) linear systems, and we give the basic evolution equation of such a system. Section 3 introduces the definitions and Section 4 studies the particular case of a FIFO (First In First Out) queue, to give an idea of the proof in the general case. The main goal of Section 5 is to show that the workload in a (max,+) linear system is multimodular with respect to the arrival sequence. In Section 6, we give the proof of multimodularity for the traveling time of a customer in the system. Finally, in Section 7, we show that a balanced admission policy is optimal among all deterministic policies with a given admission rate by applying the results established in [2].

2 Dynamics of (max,+) linear systems

In this section, we will describe the class of networks called *stochastic event graphs*, often called (max,+) linear systems since their dynamics follows a vectorial evolution equation in the so-called (max,+) algebra. More formally, an event graph is a Petri net $\mathcal{G} = (\mathcal{P}, \mathcal{Q}, \mathcal{F}, \mathcal{M}_0)$, where $\mathcal{P} = \{p_1, \cdots, p_P\}$ is the set of places, $\mathcal{Q} = \{q_1, \cdots, q_Q\}$ is the set of transitions (also called *nodes* for event graphs), $\mathcal{F} \subset \mathcal{Q} \times \mathcal{P} \cup \mathcal{P} \times \mathcal{Q}$ is the set of arcs and $\mathcal{M}_0 : \mathcal{P} \to \{0, 1, 2, \cdots M\}$ is the initial number of tokens in each place.

An event graph is a Petri net where each place has one input transition and one output transition. In a stochastic event graph, the n-th *firing* duration of transition q_i in $\mathcal{Q}$ is assumed to be a random variable σ_n^i with finite expectation.

A more detailed presentation of (stochastic) event graphs and Petri nets is available in numerous references, like for example [8, 4].

To any stochastic event graph, we can associate a set of matrices, $A(0,n), \cdots, A(M,n)$, all of size $Q \times Q$, where the entry (i,j) in matrix $A(k,n)$ is σ_{n-k}^j if there exists a place between transitions q_j and q_i with k initial tokens, and $-\infty$ otherwise.

If an event graph contains transitions with no input places (also called an *input* transition), we define the matrix B which gives the connections of the input with the regular transitions. More precisely, the entry i, j in B is 0 if there is a place between the input transition q_j and transition q_i, and is $-\infty$ otherwise. Also, the vector U_n denotes the firing epochs of all the input transitions.

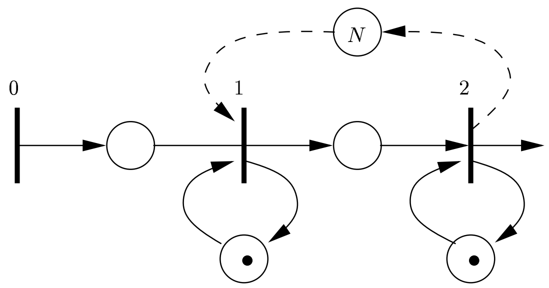

Figure 1: Event graph model of two queues in tandem

The example of two queues in tandem is given in Figure 1. If the dashed arcs are taken into account, the system corresponds to two queues in tandem with a second queue with a finite buffer of size N, and blocking before service in the first queue when the second queue is full.

In this example (with finite capacity $N = 1$), let us call as before σ_n^i the n-th firing duration of transition i and U_n (which is a scalar here) the sequence of the firing epochs of the single input transition 0, which represents the input in the system. Then we get

$$A(0,n) = \begin{pmatrix} -\infty & -\infty \\ \sigma_n^1 & -\infty \end{pmatrix}, A(1,n) = \begin{pmatrix} \sigma_{n-1}^1 & \sigma_{n-1}^2 \\ -\infty & \sigma_{n-1}^2 \end{pmatrix},$$

$$B = \begin{pmatrix} 0 \\ -\infty \end{pmatrix}.$$

Now, in the general case, if we consider the state variables $X_i(n)$ which denote the time when transition i initiates its n-th firing, then the vector $X(n) = (X_1(n), \cdots, X_Q(n))$ satisfies the following linear evolution equation in the (max,+) notation[1],

$$X(n) = A(0,n) \otimes X(n) \oplus A(1,n) \otimes X(n-1) \oplus \cdots \oplus A(M,n) \otimes X(n-M) \oplus B \otimes U_n. \tag{1}$$

[1]In the (max,+) semi-ring, the operations are $\oplus \stackrel{\text{def}}{=} \max$, $\otimes \stackrel{\text{def}}{=} +$ and for operations on matrices, if A and B are matrices of size $n \times \ell$ and $\ell \times m$ respectively, then $A \otimes B$ is a matrix of size $n \times m$ where $(A \otimes B)_{i,j} \stackrel{\text{def}}{=} \oplus_{k=1}^{\ell} A_{ik} \otimes B_{kj}$, $i = 1 \cdots n, j = 1 \cdots m$, see [4] for more details.

In the case of tandem queues, given in Figure 1, $U_n = T_n$ is the sequence of the arrival times of customers in the first queue, and the general evolution equation reduces to

$$\begin{pmatrix} X_1(n) \\ X_2(n) \end{pmatrix} = A(0,n) \otimes \begin{pmatrix} X_1(n) \\ X_2(n) \end{pmatrix} \oplus A(1,n) \otimes \begin{pmatrix} X_1(n-1) \\ X_2(n-1) \end{pmatrix} \oplus B \otimes T_n. \tag{2}$$

In the general case, after simplifications and state expansion (see Section 2.5 in [4] and [3]), one can construct a matrix denoted $A(n)$ from the matrices $\{A(i,n)\}_{i=0,\cdots M}$ such that the general implicit equation (1) can be written under its standard form:

$$X(n+1) = A(n) \otimes X(n) \oplus B(n) \otimes U_{n+1}. \tag{3}$$

As for the example, the solution of Equation (2) is

$$\begin{pmatrix} X_1(n+1) \\ X_2(n+1) \end{pmatrix} = A(n) \otimes \begin{pmatrix} X_1(n) \\ X_2(n) \end{pmatrix} \oplus B(n) \otimes (T_{n+1}). \tag{4}$$

We give the value of $A(n)$ and $B(n)$:

$$A(n) = \begin{pmatrix} \sigma_n^1 & \sigma_n^2 \\ \sigma_n^1 + \sigma_{n+1}^1 & \sigma_n^2 + \sigma_{n+1}^1 \end{pmatrix}, B(n) = \begin{pmatrix} 0 \\ \sigma_{n+1}^1 \end{pmatrix}.$$

Note that it is a general fact that the entries of $A(n)$ are functions of the $\{\sigma_n^i\}_{n\in\mathbb{N}}$ only.

3 Multimodularity and admission control

3.1 Definition of multimodularity

For more details on multimodularity and its properties, please refer to [10, 2]. Here, we only give the basic definitions.

Let $e_i \in \mathbb{N}^m$, $i = 1, ..., m$ denote the vector having all entries equal to zero except for a "1" in entry i. Define $d_i = e_i - e_{i+1}$, $i = 1, ..., m-1$. Let $\{d_0 = -e_1, d_1, ..., d_{m-1}, d_m = e_m\}$ be the basis-set of vectors for multimodularity. We have:

$$\begin{array}{rcl} d_0 & = & (\begin{array}{cccccc} -1 & 0 & 0 & \cdots & 0 & 0 \end{array}) \\ d_1 & = & (\begin{array}{cccccc} 1 & -1 & 0 & \cdots & 0 & 0 \end{array}) \\ d_2 & = & (\begin{array}{cccccc} 0 & 1 & -1 & \cdots & 0 & 0 \end{array}) \\ & & \vdots \\ d_{m-1} & = & (\begin{array}{cccccc} 0 & 0 & 0 & \cdots & 1 & -1 \end{array}) \\ d_m & = & (\begin{array}{cccccc} 0 & 0 & 0 & \cdots & 0 & 1 \end{array}). \end{array} \tag{5}$$

Following Hajek [10], we say that a function f defined on $\mathbb{Z}^m$ is *multimodular* if $f(a + d_i) + f(a + d_j) \geqslant f(a) + f(a + d_i + d_j)$ for all $i \neq j$. As noted by Hajek, this is equivalent to the fact that $f(a - d_i) + f(a - d_j) \geqslant f(a) + f(a - d_i - d_j)$.

In this paper, we will use functions which are only defined on the positive orthant, $\mathbb{N}^m$ (which is a convex subset on $\mathbb{Z}^m$) and we will only require that for all d_i and d_j with $i \neq j$ such that $a, a - d_i, a - d_j, a - d_i - d_j$ are all in $\mathbb{N}^m$, we have $f(a - d_i) + f(a - d_j) \geqslant f(a) + f(a - d_i - d_j)$.

This restriction to $\mathbb{N}^m$ does not alter the main properties of multimodular functions as shown in [2]. In particular, convexity properties are preserved by this partial definition as well as the optimization results used in Theorems 7.5 and 7.13.

However, this restriction is essential in our application of multimodularity to admission control since we have not been able to give a consistent meaning to negative customers in this context.

3.2 Admission policy: the time slot approach

In the rest of the paper, we will use the following notations:

let $\{T_i\}_{i\in\mathbb{N}}$ be the instants of arrival opportunities, with the convention that $T_1 = 0$. Since all the rest of the notations are based on the original sequence, this can be considered as a time driven approach.

We denote by δ_i the ith *interval* length, that is: $\delta_i = T_{i+1} - T_i$. We assume that $\delta_0 = 0$. From now, the sequence T is fixed (and hence the sequence δ).

As for the customers, they arrive by batches. Their arrival is defined through an *arrival sequence* which is a sequence of integer numbers, $a = (a_1, a_2, \cdots, a_N, \cdots)$, where a_i gives the number of customers entering to the queue at time T_i. For convenience, we introduce the value $a_0 = 0$. In Section 7 the sequence a will be seen as a control sequence over the arrivals but will be used in the same way.

Furthermore, each individual customer carries a load. The load of the $j-$th customer is denoted σ_j.

The *counting function* $\kappa(i)$ is the number of individual arrivals by time T_i:

$$\kappa(i) \stackrel{\text{def}}{=} \sum_{j=1}^{i} a_j. \tag{6}$$

The function $\nu(i)$ is the *number of time-intervals* elapsed when the ith individual customer enters the system. We assume that $\nu(0) = 0$ and

$$\nu(i) \stackrel{\text{def}}{=} \min\{m : \sum_{j=1}^{m} a_j \geqslant i\}. \tag{7}$$

The function $\tau_i(a)$ is the time elapsed between the i-th and $(i+1)$-th individual arrivals, defined by:

$$\tau_i(a) \quad \stackrel{\text{def}}{=} \quad \sum_{j=\nu(i)}^{\nu(i+1)-1} \delta_j, \quad i < \kappa(N), \tag{8}$$

and for $i = \kappa(N)$, we set $\tau_{\kappa(N)}(a) = T_N - T_{\nu(\kappa(N))}$.

Figure 2 illustrates all these preliminary definitions.

4 The FIFO queue

In this section, we assume that the arrival stream enters a single FIFO queue. This is a simple case and the proofs presented here are typical of what happens in the general framework.

The system considered is a G/G/1 queue with batch arrivals. Note that according to Section 2, this system is (max,+) linear. If $a_i = k$, then this means that k customers enter the queue at time k.

In this part, we show that the function $\mathbf{E}W$, the expected workload, is multimodular on $\mathbb{N}^N$. We denote by $\mathsf{W}_j(a)$, the workload in the system just before the j-th individual arrival, for $j \leqslant \kappa(N)$ and $\mathsf{W}_{\kappa(N)+1}(a)$ is the workload at time T_N under the arrival sequence (a). It satisfies the following recurrence equation.

$$\mathsf{W}_{\kappa(N)+1}(a) = (\mathsf{W}_{\kappa(N)}(a) + \sigma_{\kappa(N)} - \tau_{\kappa(N)}) \vee 0. \tag{9}$$

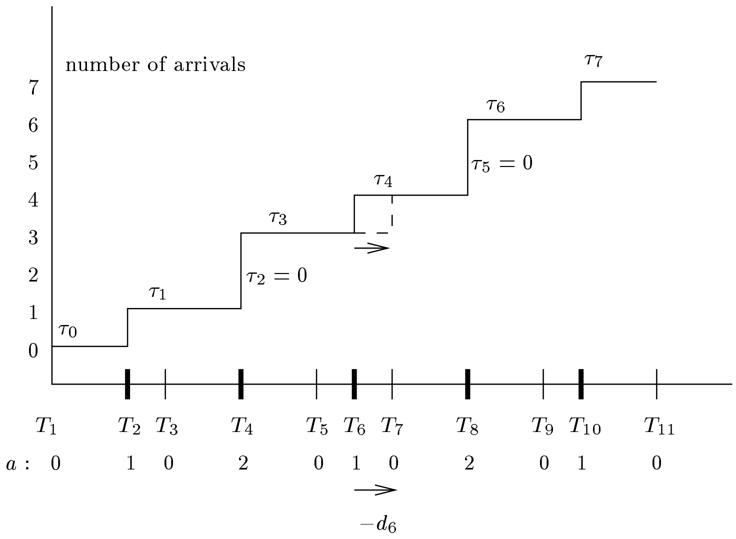

Figure 2: an example

The solution of this equation is given by the expanded Lindley's equation:

$$\mathsf{W}_{\kappa(N)+1}(a) = \max\left(0, (\sigma_{\kappa(N)} - \tau_{\kappa(N)}), (\sigma_{\kappa(N)} - \tau_{\kappa(N)}) + ((\sigma_{\kappa(N)-1} - \tau_{\kappa(N)-1}), \cdots\right). \qquad (10)$$

For convenience, we denote $W_N(a) \stackrel{\text{def}}{=} \mathsf{W}_{\kappa(N)+1}(a)$ and

$$w_j \stackrel{\text{def}}{=} \sum_{h=j}^{\kappa(N)} (\sigma_h - \tau_h).$$

Using this definition, we have:

$$W_N(a) = \max\left(0, w_{\kappa(N)}, \cdots, w_1\right).$$

4.1 Coupling of the service times with the customers

The rest of the section is devoted to proving multimodularity of the workload. Unfortunately, this does not hold on sample paths (as illustrated by the following example). We have to use a coupling of the service times with the customers entering the queue, and then, by assuming that the service times are stationary, we will prove multimodularity of the expected workload.

First, we will illustrate the difficulty of attaching the service times to customers while insuring that multimodularity holds, through an example.

An example We consider a single queue with a sequence of service times $\sigma(n) = (4, 1, 1, 1, 1, 1, \cdots)$ and with the integer points as arrival epochs.

We focus on the workload at time $T = 7$ under the arrival streams:

$$\begin{aligned} a &= (0,0,1,0,0,1,1,0) && (11)\\ a - d_0 &= (1,0,1,0,0,1,1,0) && (12)\\ a - d_6 &= (0,0,1,0,0,0,2,0) && (13)\\ a - d_6 - d_0 &= (1,0,1,0,0,0,2,0) && (14) \end{aligned}$$

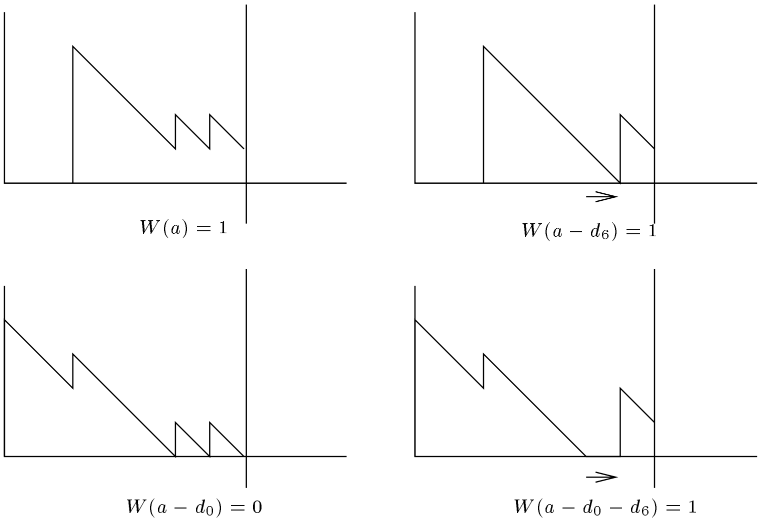

Figure 3: The workload is not multimodular.

The workload satisfies: $W_7(a) = 1, W_7(a - d_0) = 0, W_7(a - d_0 - d_6) = 1, W_7(a - d_6) = 1$, as shown in Figure 3.

This shows that the function W_7 is not multimodular, since $W_7(a) + W_7(a - d_0 - d_6) > W_7(a - d_0) + W_7(a - d_6)$.

However, under a proper coupling of the service times with the customers, then W can be made multimodular.

We couple the service times with the customers entering the queue in the following way.

$$
\begin{array}{cccccccccc}
a & = & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
\sigma & = & & & \sigma_1 & & & \sigma_2 & \sigma_3 & \\
\\
a - d_0 & = & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
\sigma & = & \sigma_0 & & \sigma_1 & & & \sigma_2 & \sigma_3 & \\
\\
a - d_6 & = & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 \\
\sigma & = & & & \sigma_1 & & & & \sigma_2, \sigma_3 & \\
\\
a - d_0 - d_6 & = & 1 & 0 & 1 & 0 & 0 & 0 & 2 & 0 \\
\sigma & = & \sigma_0 & & \sigma_1 & & & & \sigma_2, \sigma_3 &
\end{array}
$$

Under this coupling c, the workloads become as in Figure 4. We see that $W_7^c(a) = 0, W_7^c(a - d_0) = 0, W_7^c(a - d_0 - d_6) = 1, W_7^c(a - d_6) = 1$, and W_7^c satisfies the multimodular inequality (which is an equality here),

General coupling In general, the coupling of the service times with the arrival stream is done in the following way.

- Let a be an arbitrary arrival sequence in $\mathbb{N}^N$: $a = (a_1, a_2, \cdots, a_N)$. The service times are coupled with the customers entering the queue in the following fashion: With arriving batch a_i, we attach the service times:

$$\sigma_{\kappa(i-1)+1}, \sigma_{\kappa(i-1)+2}, \cdots, \sigma_{\kappa(i)}.$$

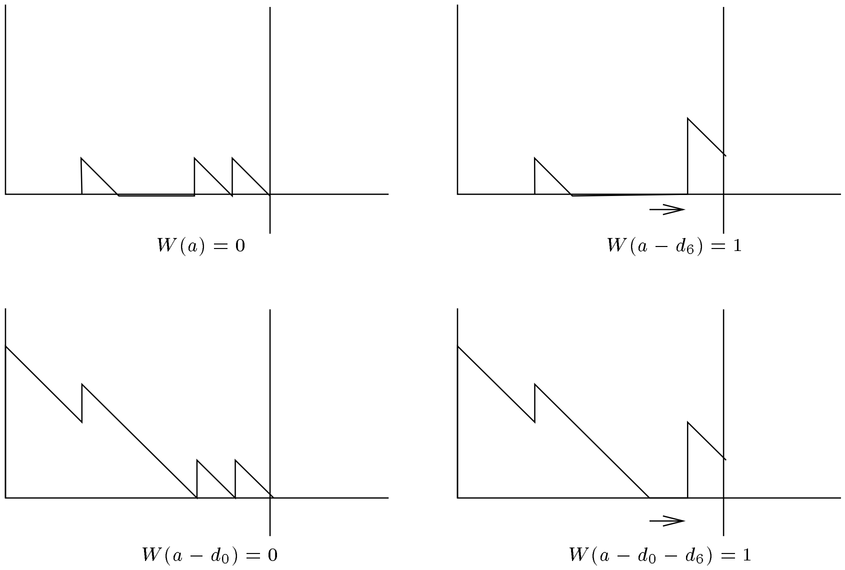

Figure 4: Multimodularity of the workload under proper coupling.

- With arrival sequence, $a - d_0$, the service time sequence for batch $a_i, i \neq 1$, is not modified and the batch attached with a_1 becomes:

$$\sigma_0, \sigma_1, \sigma_2, \cdots, \sigma_{\kappa(1)},$$

where σ_0 is a new service time.

- If $a - d_j \in \mathbb{N}^N$, $1 \leqslant j < N$, the service time sequence for batch $a_i, i \neq j, i \neq j+1$, is not modified and the batch attached with a_j (which is not empty) becomes:

$$\sigma_{\kappa(j-1)+1}, \sigma_{\kappa(j-1)+2}, \cdots, \sigma_{\kappa(j)-1}$$

and the batch attached with a_{j+1} becomes:

$$\sigma_{\kappa(j)}, \sigma_{\kappa(j)+1}, \cdots, \sigma_{\kappa(j+1)}.$$

In other words, the service time $\sigma_{\kappa(j)}$ is moved from the j-th batch to the $j-1$-th batch.

- If $a - d_N \in \mathbb{N}^N$, the service time sequence for batch $a_i, i \neq N$, is not modified and the batch attached with a_N (which was not empty) becomes:

$$\sigma_{\kappa(N-1)+1}, \sigma_{\kappa(N-1)+2}, \cdots, \sigma_{\kappa(N)-1}.$$

The last service time $\sigma_{\kappa(N)}$ has been removed.

- If $a - d_j - d_i \in \mathbb{N}^N$, the coupling of the service times is obtained by composing the modification of the coupling associated with a induced by d_i and d_j. This construction is commutative.

Note that this coupling strongly depends on the initial choice of a. If one changes the starting point a, then other service times will be chosen. Also, the coupling is not defined on a's which do not lay in the positive orthant, $\mathbb{N}^N$.

4.2 Multimodularity

We choose a point a in $\mathbb{N}^N$ and we construct the associated coupling c. We will denote by $W_N^c(.)$ the workload at time T_N under this coupling.

Now, using the notations given in the preliminaries, we have the following results.

Property 4.1. *Let* $1 \leqslant i < N$. *If* $a_i > 0$, $\tau_{\kappa(i)-1}(a-d_i) = \tau_{\kappa(i)-1}(a) + \delta_i$, $\tau_{\kappa(i)}(a-d_i) = \tau_{\kappa(i)}(a) - \delta_i$. *All other* τ_j *are unchanged.*

Proof. The addition of $-d_i$ corresponds to delaying the last acceptance at time T_i to time T_{i+1}. $\tau_{\kappa(i)}$ is the time interval between T_i and the arrival instant of the next customer that arrives after time T_i. Hence, delaying the arrival from T_i to time T_{i+1} results in increasing $\tau_{\kappa(i)-1}$ by $T_{i+1} - T_i = \delta_i$. $\tau_{\kappa(i)}$ decreases by that value. See Figure 2 for an illustration of this proof. □

Property 4.2. *Let* $0 < i < N$. *If* $a_i > 0$, $w_{\kappa(i)}(a-d_i) = w_{\kappa(i)} + \delta_i$. *All others* w_j *are left unchanged.*

Proof. Follows from Property 4.1 and the definition of w_j. □

Under the previous coupling, we have the following results.

Lemma 4.3. *Let* $h : \mathbb{R} \to \mathbb{R}$ *be a nondecreasing convex function. Then* $h \circ W_N^c(a) + h \circ W_N^c(a - d_i - d_j) \leqslant h \circ W_N^c(a-d_i) + h \circ W_N^c(a-d_j)$, *for all* i, j *and all* a *such that* $a - d_i - d_j$, $a - d_i$ *and* $a - d_j$ *are in* $\mathbb{N}^N$.

Proof.

- Let us first consider the case $0 < i, j < N$ and $a_i > 0, a_j > 0$. In this case, using the previous Property 4.2, we have:

$$
\begin{aligned}
W_N^c(a-d_i) &= \max(W_N^c(a), w_{\kappa(i)} + \delta_i) && (15)\\
W_N^c(a-d_j) &= \max(W_N^c(a), w_{\kappa(j)} + \delta_j) && (16)\\
W_N^c(a-d_i-d_j) &= \max(W_N^c(a), w_{\kappa(i)} + \delta_i, w_{\kappa(j)} + \delta_j). && (17)
\end{aligned}
$$

  Therefore, we have

$$
\begin{aligned}
\max(W_N^c(a), W_N^c(a-d_i-d_j)) &= \max(W_N^c(a-d_i), W_N^c(a-d_j)), && (18)\\
W_N^c(a) + W_N^c(a-d_i-d_j) &\leqslant W_N^c(a-d_i) + W_N^c(a-d_j). && (19)
\end{aligned}
$$

  This means that the two dimensional vector $(W_N^c(a), W_N^c(a-d_i-d_j))$ is weakly majorized by $(W_N^c(a-d_i), W_N^c(a-d_j))$. By using Theorem B.2, p.109 in [11], for any non-decreasing convex function h, $h(W_N^c(a)) + h(W_N^c(a-d_i-d_j)) \leqslant h(W_N^c(a-d_i)) + h(W_N^c(a-d_j))$.

- Let us now assume that $i = 0$. This case corresponds to the arrival of an extra customer at time $T_1 = 0$. This customer brings a load that we denote σ_0. In this case, we have:

$$W_N^c(a-d_0) = \max(W_N^c(a), w_1 + (\sigma_0 - \tau_0)). \qquad (20)$$

  This case is treated similarly to the case above since $W_N^c(a-d_0)$ is of the form $\max(W_N^c(a), X)$, where $X = w_1 + (\sigma_0 - \tau_0))$.

- If $i = N$ and $a_N > 0$, then we have:

$$W_N^c(a-d_N) = (W_N^c(a) - \sigma_{\kappa(N)})^+. \qquad (21)$$

  Here, for some X (see the first two cases)

$$W_N^c(a-d_j) = \max(W_N^c(a), X), \qquad (22)$$

and

$$W_N^c(a-d_j-d_N) = (\max(W_N^c(a), X) - \sigma_{\kappa(N)})^+. \tag{23}$$

By a case analysis, we see that if $W_N^c(a-d_j-d_N) = 0$ then, $W_N^c(a-d_N) = 0$ and by monotonicity of h, $h \circ W_N^c(a-d_N) + h \circ W_N^c(a-d_j) \geqslant h \circ W_N^c(a) + h \circ W_N^c(a-d_j-d_N)$. If $W_N^c(a-d_j-d_N) > 0$, then, $W_N^c(a-d_j) - W_N^c(a-d_j-d_N) = \sigma_{\kappa(N)}$. This yields,

$$W_N^c(a) + W_N^c(a-d_j-d_N) \leqslant W_N^c(a-d_N) + W_N^c(a-d_j).$$

On the other hand, it is direct to see that

$$\max(W_N^c(a), W_N^c(a-d_j-d_N)) \leqslant \max(W_N^c(a-d_N) + W_N^c(a-d_j)).$$

Again, by using the same Schur convexity property, for h convex and increasing,

$$h \circ W_N^c(a) + h \circ W_N^c(a-d_j-d_N) \leqslant h \circ W_N^c(a-d_N) + h \circ W_N^c(a-d_j).$$

This concludes the proof. □

Theorem 4.4. *Suppose $\{\sigma_n\}_{n=1}^{\infty}$ is a stationary sequence, then the function $\mathbf{E}_\sigma(h \circ W_N)$ is multimodular, where $\mathbf{E}_\sigma$ denotes the expectation w.r.t. the sequence $\{\sigma_n\}$.*

Proof. Let a be an arbitrary point in $\mathbb{N}^N$. Construct the associated coupling c of the services times. Under this coupling and for all i, j, Lemma 4.3 shows that $h \circ W_N^c(a-d_i) + h \circ W_N^c(a-d_j) - h \circ W_N^c(a) - h \circ W_N^c(a-d_i-d_j) \geqslant 0$. Therefore, $\mathbf{E}_\sigma\Big(h \circ W_N^c(a-d_i) + h \circ W_N^c(a-d_j) - h \circ W_N^c(a) - h \circ W_N^c(a-d_i-d_j)\Big) \geqslant 0$. By the stationary assumption on the service times and using the fact that under coupling c, the service times involved in $W_N^c(.)$ are always consecutive, $\mathbf{E}_\sigma h \circ W_N^c(.) = \mathbf{E}_\sigma h \circ W_N(.)$, for the points $a-d_i$, $a-d_i$, $a-d_i-d_j$ and a since the expectation is invariant with respect to the shift operator. Finally we get
$\mathbf{E}_\sigma h \circ W_N(a-d_i) + \mathbf{E}_\sigma h \circ W_N(a-d_j) - \mathbf{E}_\sigma h \circ W_N(a) - \mathbf{E}_\sigma h \circ W_N(a-d_i-d_j) \geqslant 0$. □

5 The general case: (max,+) systems with one input

The G/G/1 queue is a special case of a (max,+) linear system. This section will generalize the multimodularity properties to the case of an arbitrary network which is (max,+) linear and has one input.

5.1 Description of the system

In this section, we will consider a general stochastic event graph with a single input transition. As mentioned in Section 2, the dynamic behavior of this system is linear in the (max,+) algebra.

Let $\mathcal{G} = (\mathcal{P}, \mathcal{Q}, \mathcal{F}, \mathcal{M}_0)$ be an event graph with Q nodes (or transitions) and only one input (which can be seen as a source of exogenous costumers) and which is *input-connected* (*i.e.* for each transition q in the net, there exists a path from the input to q). The single input node, is represented by a special transition q_0 in the marked graph.

For each transition (or node) q in $\mathcal{G}$, we consider all paths π from q_0 to q. This set is denoted by $\mathcal{P}(q)$. We also denote by $M(\pi) \stackrel{\text{def}}{=} \sum_{s\in\pi} \mathcal{M}_0(s)$, the sum of the initial tokens on the path π. Now, we define

$$L(q) = \min_{\pi \in \mathcal{P}(q)} M(\pi).$$

Lemma 5.1. *When an event graph complies with the foregoing assumptions, the $n+L(q)$-th firing of transition q of $\mathcal{G}$ involves a token produced by the n-th firing of transition q_0.*

Proof. Let s_q be the shortest path from q_0 to q with $L(q)$ tokens. The length of s_q is called the distance from q_0 to q. The proof holds by induction on the length of s_q. If $s_q = 0$, then $q = q_0$ and the result is true. Suppose that the result is true for all transitions at "distance" $k-1$ from q_0. Choose q at distance k, then the transition q' preceding q on the path s_q is at distance $k-1$ from q_0 and induction applies to q'. Now the place (q', q) contains m tokens. By definition of q', $L(q') = L(q) - m$, and by induction, the $n + L(q')$-th firing of transition q' uses the token number n and since the buffer place between q' and q is FIFO, the $n + L(q)$-th firing of q will use that token. □

In the rest of the paper, we will also assume that the system $\mathcal{G}$ is initially empty, *i.e.* for every node q in $\mathcal{G}$, there exists a path from q_0 to q that contains no tokens. More formally, this can be written,

$$L(q) = 0, \forall q \in S.$$

This assumption will be necessary to prove Lemmas 7.4 and 7.6. Note however, that the system being empty does not mean that the initial marking is zero in all the places. For instance, the network depicted in Figure 1 is empty, since $L_1 = L_2 = 0$. However, some places (as the recycling places or the place with dashed arrows) have a positive marking.

5.2 Lindley's equation for (max,+) systems

We recall that $X_q(n)$ is the epoch when transition q fires for the n-th time with the convention that $X_q(n) = 0$ if $n \leqslant 0$. Note that because of Lemma 5.1, $X_q(n)$ is also the epoch when the n-th token generated by the input is fired by transition q.

Now, let W_n be a vector , with its q-th component equal to: $W_n^q \stackrel{\text{def}}{=} X_q(n) - X_0(n)$. Using Lemma 5.1, W_n^q can be seen as the *traveling time* for customer n between its entrance in the system and its passage in transition q.

In the (max,+) algebra, we recall that we have the following general vectorial equation that describes the evolution of $\mathcal{G}$ (see Equation (3) and Section 2.5 in [4] or [6] for examples):

$$X(n+1) = A(n) \otimes X(n) \oplus B(n) \otimes U_{n+1}.$$

If we consider only the component $X_i(n)$ and the row i in matrix $A(n)$ (denoted $A_i(n)$), we can subtract $U_{n+1} = T_{n+1} = X_0(n+1)$ on each side of this equation and we get:

$$W_{n+1}^i = A_i(n) \otimes X(n) \oplus B(n)_i \otimes T_{n+1} - T_{n+1}.$$

We can rewrite this as

$$\begin{aligned} W_{n+1}^i &= A_i(n) \otimes (X(n) - T_{n+1}) \oplus B(n)_i \\ &= (A_i(n) - \tau_n) \otimes (X(n) - T_n) \oplus B(n)_i \end{aligned}$$

If we write this last equality for $W(n)$ in vectorial form, we get

$$W_{n+1} = A(n) \otimes D(-\tau_n) \otimes W_n \oplus B(n),$$

where $D(h)$ is the diagonal matrix with h on the diagonal and $-\infty$ everywhere else.

This recursion is a generalization of the Lindley equation in the case of a network. This equation was also derived in [6].

By using elementary matrix operations in the (max,+) algebra, the equation can also be developed into:

$$W_{n+1} = B(n) \oplus \bigoplus_{i=1}^n C_i \qquad ,$$
$$\text{with } C_i = \bigotimes_{j=i}^n \big(A(j) \otimes D(-\tau_j)\big) \otimes B(n-i-1), \text{ and } B(0) = (-\infty, \cdots, -\infty)^t.$$

5.3 Multimodularity

The main result established in this section is that the expectation of W_n^q is multimodular for all q. The proof is very similar to the case of the single queue and is made, surprisingly, even easier by using the vectorial form of the Lindley equation in the (max,+) algebra.

In the following, we will also often use the following transformation for notational convenience. If X is a vector of size Q, then bold face $\mathbf{X}$ is a diagonal matrix of size $Q \times Q$, with the vector X on the diagonal and $-\infty$ elsewhere.

The multidimensional coupling of the service times in each transition with the arriving customers is done similarly as in the one queue case.

We construct a coupling for transition q which is independent of the coupling for any other transition.

- Let a be an arbitrary arrival sequence in $\mathbb{N}^N$: $a = (a_1, a_2, \cdots, a_N)$. The service times are coupled with the customers entering the system in the following fashion:

  With arriving batch a_i, we attach the service times: $\sigma^q_{\kappa(i-1)+1}, \sigma^q_{\kappa(i-1)+2}, \cdots, \sigma^q_{\kappa(i)}$.
- With arrival sequence, $a - d_0$, the service time sequence for batch $a_i, i \neq 0$ is not modified and the batch attached to a_1 now becomes under this coupling: $\sigma^q_0, \sigma^q_1, \sigma^q_2, \cdots, \sigma^q_{\kappa(1)}$ where σ^q_0 is a new service time.
- If $a - d_j \in \mathbb{N}^N$, then with arrival sequence, $a - d_j$, $1 < j < N$ the service time sequence for batch $a_i, i \neq j, j-1$ is not modified, the batch attached to a_{j-1} becomes: $\sigma^q_{\kappa(j-2)+1}, \sigma^q_{\kappa(j-2)+2}, \cdots, \sigma^q_{\kappa(j-1)-1}$ and finally the batch attached to a_j also becomes: $\sigma^q_{\kappa(j-1)}, \sigma^q_{\kappa(j-1)+1}, \cdots, \sigma^q_{\kappa(j)}$. In other words, the service time $\sigma^q_{\kappa(j-1)}$ is moved from the $(j-1)$-th batch to the j-th batch.
- If $a - d_N \in \mathbb{N}^N$, then with arrival sequence, $a - d_N$, the service time sequence for batch $a_i, i \neq N$ is not modified and the batch attached to a_N (which is not empty) becomes: $\sigma^q_{\kappa(N-1)+1}, \sigma^q_{\kappa(N-1)+2}, \cdots, \sigma^q_{\kappa(N)-1}$. The last service time $\sigma^q_{\kappa(N)}$ has been removed.
- If $a - d_j - d_i \in \mathbb{N}^N$, then with arrival sequence, $a - d_j - d_i$, the coupling of the service times is obtained by composing the modification of the coupling associated with a induced by d_i and d_j. We note again that the construction is commutative.

Now, using this coupling, we consider the traveling time vector of a potential customer entering the system at time T_N under arriving stream a and its associated coupling. We denote this vector by $W_N(a)$. This vector is defined by the following equation:

$$W_N(a) = B(\kappa(N)) \oplus \bigoplus_{i=1}^{\kappa(N)} C_i(a), \tag{24}$$

with

$$C_i(a) = \bigotimes_{j=i}^{\kappa(N)} \big(A(j) \otimes D(-\tau_j(a))\big) \otimes B(\kappa(N) - i - 1),$$

where $\tau_j(a)$ is defined in Equation (8).

We have similar lemmas as in the one queue case, (see Properties 4.1 - 4.2).

Lemma 5.2. *Let $0 < i < N$. If $a_{i-1} > 0$, then $C_{\kappa(i)}(a - d_i) = D(\delta_i) \otimes C_{\kappa(i)}(a)$. All other C_j are left unchanged.*

Proof. Using property 4.1, we have: $D(-\tau_{\kappa(i)-1}(a-d_i)) = D(-\delta_i) \otimes D(-\tau_{\kappa(i)-1}(a))$ and $D(-\tau_{\kappa(i)}(a - d_i)) = D(\delta_i) \otimes D(-\tau_{\kappa(i)}(a))$ with all others τ_j left unchanged.

- Now, for every $j > \kappa(i)$, $C_j(a-d_i)$ does not involve $D(-\tau_{\kappa(i)}(a-d_i))$ or $D(-\tau_{\kappa(i)-1}(a-d_i))$, and therefore is left unchanged.
- If $j < \kappa(i)$, then $C_j(a-d_i)$ involves $D(-\tau_{\kappa(i)}(a-d_i))$ and $D(-\tau_{\kappa(i)-1}(a-d_i))$. Since the matrices $D(x)$ commute with everything, and since $D(\delta_i) \otimes D(-\delta_i) = E$, the identity matrix, then, $C_j(a-d_i)$ is left unchanged.
- Finally, if $j = \kappa(i)$, then $C_j(a-d_i)$ involves $D(-\tau_{\kappa(i)}(a-d_i))$ but not $D(-\tau_{\kappa(i)-1}(a-d_i))$. Using the fact that $D(\delta_i)$ commutes with all the other matrices, we have $C_{\kappa(i)}(a-d_i) = D(\delta_i) \otimes C_{\kappa(i)}(a)$.

□

In the following we will use, to simplify the equations, the matrices: $Z_i \stackrel{\text{def}}{=} D(\delta_i) \otimes C_{\kappa(i)}(a)$.

Lemma 5.3. *Let $h = (h_1, \cdots, h_Q)$ be such that for all q, $h_q : \mathbb{R} \to \mathbb{R}$ is an increasing convex function. Component-wise, we have*

$$h_q \circ W_N^q(a-d_i) \otimes h_q \circ W_N^q(a-d_j) \geqslant h_q \circ W_N^q(a) \otimes h_q \circ W_N^q(a-d_j-d_i). \tag{25}$$

Proof. First note as a general remark that for any matrix M and any positive number x, $D(x) \otimes M \oplus M = D(x) \otimes M$. Now, as in the case of a single queue, we have to distinguish three cases.

- The case where $0 < i, j < N$ and $a_i, a_j > 0$. By commutativity of the $\oplus$ operator, we have

$$\begin{aligned} W_N(a-d_i) &= W(a) \oplus Z_i & (26)\\ W_N(a-d_j) &= W(a) \oplus Z_j & (27)\\ W_N(a-d_j-d_i) &= W(a) \oplus Z_j \oplus Z_i. & (28) \end{aligned}$$

  Now using the distributivity of $\otimes$ w.r.t. $\oplus$, we have $\mathbf{W_N}(a-d_i) \otimes \mathbf{W_N}(a-d_j) = \mathbf{W_N}(a) \otimes \mathbf{W_N}(a-d_j-d_i) \oplus \mathbf{Z}_j \otimes \mathbf{Z}_i$. This last equation interpreted in the classical algebra says that for each server q, the traveling time at time T_N satisfies:

$$W_N^q(a-d_i) + W_N^q(a-d_j) \geqslant W_N^q(a) + W_N^q(a-d_j-d_i). \tag{29}$$

  The fact that h is increasing and convex component-wise, and using a case analysis with Equation (29) shows that $h \circ W_N^q(a-d_i) + h \circ W_N^q(a-d_j) \geqslant h \circ W_N^q(a) + h \circ W_N^q(a-d_j-d_i)$.

- Now we examine the case where $i = 0$. As in the single queue case, this corresponds under our coupling to the arrival of an extra customer at time $T_1 = 0$ that has a service time that we denote σ_0^q in queue q. In this case, we have with $n = \kappa(N)$:

$$W_N(a-d_0) = W_N(a) \oplus \bigotimes_{j=0}^{n} \big(A_j \otimes D(-\tau_j)\big) \otimes B_i(n). \tag{30}$$

  This case is treated as the general case, since $W_N(a-d_i)$ is of the form $W_N(a) \oplus Z$, for some vector $Z \stackrel{\text{def}}{=} \bigotimes_{j=0}^{n} \big(A_j \otimes D(-\tau_j)\big) \otimes B_i(n)$.

- If $i = N$ and $a_N > 0$. This case corresponds to the removal of the last customer in the arrival batch, which happens to have arrived at time T_N. In this case, we have:

$$W_N(a-d_N) = S \otimes W_N(a) \oplus O,$$

where O is a vector composed of zeros and S is a diagonal matrix with $S_{[q,q]} = -\sigma^q_{\kappa(N)}$. Since we also have $W_N(a - d_j) = W_N(a) \oplus Z$ and $W_N(a - d_j - d_N) = S \otimes (W_N(a) \oplus Z) \oplus O = S \otimes W_N(a) \oplus S \otimes Z \oplus O$, with $Z \stackrel{\text{def}}{=} Z_j$, we get using distributivity of $\otimes$ with respect to $\oplus$,

$$\begin{aligned}
&\mathbf{W_N}(a - d_N) \otimes \mathbf{W_N}(a - d_j) \\
&= (S \otimes \mathbf{W_N}(a) \oplus \mathbf{O}) \otimes (\mathbf{W_N}(a) \oplus \mathbf{Z}) \\
&= S \otimes \mathbf{W_N}(a) \otimes \mathbf{W_N}(a) \oplus \mathbf{O} \otimes \mathbf{W_N}(a) \oplus S \otimes \mathbf{W_N}(a) \otimes \mathbf{Z} \oplus \mathbf{O} \otimes \mathbf{Z},
\end{aligned}$$

and on the other hand,

$$\mathbf{W_N}(a) \otimes \mathbf{W_N}(a - d_j - d_N) \quad = \quad \mathbf{W_N}(a) \otimes S \otimes \mathbf{W_N}(a) \oplus \mathbf{W_N}(a) \otimes S \otimes \mathbf{Z} \oplus \mathbf{W_N}(a) \otimes \mathbf{O}.$$

Since all the matrices involved in these equations are diagonal, they commute and we have:

$$\mathbf{W_N}(a - d_N) \otimes \mathbf{W_N}(a - d_j) = \mathbf{W_N}(a) \otimes \mathbf{W_N}(a - d_j - d_N) \oplus \mathbf{O} \otimes \mathbf{Z}.$$

Since $\mathbf{O} \otimes \mathbf{Z}$ is a non-negative diagonal matrix, the result is established by rewriting this equation in the conventional algebra.

As with the function h, the proof is similar to the case of a single queue. □

Let $\sigma_n, n \geqslant 1$ be the stochastic vectors with components $\sigma^q_n, q \in \mathcal{Q}$.

Theorem 5.4. *Suppose $\{\sigma_n\}$ is a stationary sequence of stochastic vectors. Then, the function $\mathbf{E}_\sigma h(W^q_N(a))$ is multimodular, where $\mathbf{E}_\sigma$ denote the expectation w.r.t. the service times in all the nodes of the system.*

Proof. The proof is similarly to the one queue case. The coupling c is compatible with the shift of stochastic vectors of the service times in all nodes. Therefore, the inequality given in Equation (25) implies the multimodularity of the expected traveling time w.r.t. all service times. □

Note that this theorem proves that the traveling time for a customer arriving at time T_N, that is, the time between its entrance in the system and its service in queue q, is multimodular, for all q and N.

Also note that the case of event graphs with multiple or no entries as well as the case where the marked graph is not empty initially are intractable by this means. Indeed, the coupling of the service times to the customers is not feasible *a priori* in those cases. It will depend on the sequence of arrivals in the case of multiple entries, and on all the service times in the case of a closed system.

6 A dual policy: counting variable and waiting time

In the previous section, we were interested in the study of the workload which is a criterion related to the server. Here we will focus on performances related to the customers, namely, the waiting time of the customers entering the network. This approach will be dual (in some particular sense) to the previous one and may be a more important issue for practical applications, where the customer satisfaction is more important than resource optimization.

Previously, all quantities were indexed by n, the number of time slots. In this section, all quantities will rather be indexed by $\kappa(n)$, the number of arrivals. The counting sequence b, will be given in the following way: $b_n = \nu(n) - \nu(n-1)$, $n > 1$. For example if $a = (0, 0, 1, 0, 0, 2, 0, 1, 1, 0, 0, 1)$, then $b = (3, 0, 2, 1, 3)$. Note that a and b represent the same information (up to the initial arrival); a_n gives the number of arrivals at time slot n (this is a time driven concept), and the dual variable b_k gives the number of time slots elapsed between the $(k-1)$-*th* arrival and the k-*th* arrival (this is a event driven concept).

6.1 Waiting time

Let $\mathcal{G}$ be a (max,+) linear system with a single input satisfying the assumptions given in Section 5.1.

Then, the traveling time of the k-th admitted customer to node i is denoted by $\mathcal{W}_k^i(b)$. The vector $\mathcal{W}_k(b)$ satisfies the vectorial Lindley equation, using the function τ defined in equation (8):

$$\mathcal{W}_{k+1}(b) = H(k) \otimes D(-\tau_k) \otimes \mathcal{W}_k(b) \oplus B(k),$$

This can be written:

$$\mathcal{W}_{k+1} = B(k) \oplus \bigoplus_{i=1}^{n} \mathcal{C}_i, \quad \text{with} \quad \mathcal{C}_i = \bigotimes_{j=i}^{k} \big(H(j) \otimes D(-\tau_j)\big) \otimes B(k-i-1.$$

This equation has essentially the same form as the equation (24).

6.2 Coupling

The coupling adapted in this case is essentially the dual of the coupling on the service times used previously. Here we rather couple the inter-arrival times, δ_i. This coupling uses the function ν instead of the function κ.

We build the coupling d in the following way.

- Let b be an arbitrary arrival sequence in $\mathbb{N}^N$: $b = (b_1, b_2, \cdots, b_N)$. The intervals are coupled with the customers entering the queue in the following fashion: With interval length b_i, we attach the intervals: $\delta_{\nu(i-1)+1}, \delta_{\nu(i-1)+2}, \cdots, \delta_{\nu(i)}$.
- With $b+d_0$, the interval sequence for length $b_i, i \neq 1$ is not modified and the length attached with b_1 (which was not empty) becomes: $\delta_1, \delta_2, \cdots, \delta_{\nu(1)}$, where δ_0 has been removed.
- If $b + d_j \in \mathbb{N}^N$, then with arrival sequence, $b + d_j$, $1 < j < N$ the interval sequence for length $b_i, i \neq j, j-1$ is not modified and the length attached with b_{j-1} becomes: $\delta_{\nu(j-2)+1}, \delta_{\nu(j-1)+2}, \cdots, \delta_{\nu(j-1)}, \delta_{\nu(j-1)+1}$ and the length attached with b_j (which is not empty) becomes: $\delta_{\nu(j-1)+2}, \cdots, \delta_{\nu(j)}$. In other words, the interval $\delta_{\nu(j-1)+1}$ is moved from the j-th length to the $(j-1)$-*th* length.
- If $b + d_N \in \mathbb{N}^N$, then with arrival sequence, $b + d_{N+1}$, the interval sequence for length $b_i, i \neq N$ is not modified and the length attached with b_N becomes $\delta_{\nu(N-1)+1}, \delta_{\nu(N-1)+2}, \cdots, \delta_{\nu(N)-1}, \delta_{\nu(N)+1}$. The last interval $\delta_{\nu(N)+1}$, is a new interval.
- If $b+d_j+d_i \in \mathbb{N}^N$, then with arrival sequence, $b+d_j+d_i$, the coupling of the δ's is obtained by composing the modification of the coupling associated with b induced by d_i and d_j. This construction is commutative.

6.3 Multimodularity

In this case, we will use the direct base $\{d_0, \cdots, d_N\}$ rather than $\{-d_0, \cdots, -d_N\}$ used in the previous sections.

Property 6.1. *Let $0 < i < N$. If $b_i > 0$, then $\tau_i(b + d_i) = \tau_i(b) + \delta_{\nu(i)+1}$ and $\tau_{i+1}(b + d_i) = \tau_{i+1}(b) - \delta_{\nu(i)+1}$. All others τ_j are unchanged.*

Proof. By definition. See Figure 2 □

Lemma 6.2. *Let $0 < i < N$. If $b_i > 0$, then $\mathcal{C}_{i+1}(b+d_i) = D(\delta_{\nu(i)+1}) \otimes C_{i+1}(b)$. All other $\mathcal{C}_j$ are left unchanged.*

Proof. The proof is similar to the proof of Lemma 5.2 and follows from the Property 6.1 and the definition of $\mathcal{C}_j$ □

In the following we will simplify the equations using the variables: $\mathcal{Z}_i \stackrel{\text{def}}{=} D(\delta_{\nu(i)+1}) \otimes \mathcal{C}_{i+1}(b)$. Under the previous coupling, we have the following results:

Lemma 6.3. *Let $h : \mathbb{R}^Q \to \mathbb{R}^Q$ be a increasing convex function. Then, $h \circ \mathcal{W}(b) + h \circ \mathcal{W}(b + d_i + d_j) \leqslant h \circ \mathcal{W}(b + d_i) + h \circ \mathcal{W}(b - d_j)$, for all i, j and all b such that $b + d_i + d_j$, $b + d_i$ and $b + d_j$ are in $\mathbb{N}^N$.*

Proof. The proof is essentially similar to the proof of lemma 5.3. The proof follows from the following equalities by using the same technique as in the previous case.

- If $0 < i, j < N$,

$$\begin{aligned} \mathcal{W}(b + d_i) &= \mathcal{W}(b) \oplus \mathcal{Z}_i & (31)\\ \mathcal{W}(b + d_j) &= \mathcal{W}(b) \oplus \mathcal{Z}_j & (32)\\ \mathcal{W}(b + d_j + d_i) &= \mathcal{W}(b) \oplus \mathcal{Z}_j \oplus \mathcal{Z}_i. & (33) \end{aligned}$$

If $i = 0$,

$$\mathcal{W}(b + d_0) = \mathcal{W}(b) \oplus \mathcal{Z}_0. \tag{34}$$

If $i = N$,

$$\mathcal{W}(b + d_N) = \mathcal{S} \otimes \mathcal{W}(b) \oplus \mathcal{O}, \tag{35}$$

where $\mathcal{O}$ is a vector composed of zeros and $\mathcal{S}$ is a diagonal matrix with $\mathcal{S}_{[q,q]} = -\delta^q_{\nu(N)+1}$. □

Theorem 6.4. *If the intervals $\{\delta_i\}_{i=0}^{\infty}$ form a stationary sequence, then the function $\mathbf{E}_\delta(h \circ \mathcal{W})$ is multimodular.*

The proof is essentially similar to the proof of Theorem 5.4.

7 Optimal admission sequence

In this section, we study the admission control in a (max,+) linear system. We use the multimodularity property to derive the optimality of the most regular admission sequence.

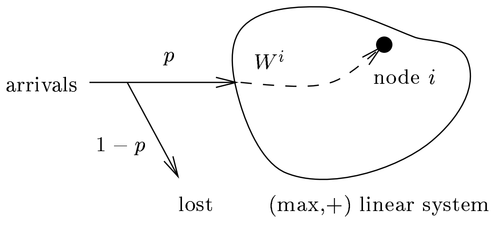

Figure 5: Illustration of the control problem

We want to admit customers in a (max,+) linear system $\mathcal{G}$, under the constraint that in the long run, the fraction of customers lost be at most $1 - p$.

Now the problem is to find which admission policy minimizes the traveling time to node i of a customer admitted in the system, as illustrated in the figure 5. The admission is governed by a binary control sequence $a = (a_1, a_2, \cdots)$. If $a_n = 1$ then the *n-th* customer is admitted to the queue. If $a_n = 0$ then the *n-th* customer is lost.

Note that this *control sequence* can be seen as an *arrival sequence* as used in the previous sections.

In this model, the vectors $\sigma_n, n = 1, 2, \cdots$ are service times of the customers which are admitted to the system, they correspond precisely to the definition of the firing times introduced in Section 2. There are no service times attached to the customers who are not admitted to the queue. This framework is natural in the stochastic Petri net context (see for example [3]). Note also that if the service times (or firing times) are independent, then this assumption is not needed and service times can as well be attached to all customers.

7.1 Balanced sequences

We first present the notion of balanced sequence that will give the optimal admission sequence. More details on these notions are presented in [1] and references therein.

Definition 7.1. *The sequence a is balanced if and only if for any d, j, l,*

$$\left| \sum_{i=j}^{j+d} a_i - \sum_{i=l}^{l+d} a_i \right| \leqslant 1$$

The following theorem is proved by Morse and Hedlund in [12] using methods from combinatorics.

Theorem 7.2. *Let a be the sequence $a = (a_1, a_2, \cdots)$. (i)- If there exists two positive real numbers θ and p, such that for all n $a_n = \lfloor np + \theta \rfloor - \lfloor (n-1)p + \theta \rfloor$,* [2] *then a is balanced. The quantity θ is the initial phase and p is the rate of a. (ii)- Conversely, if a is balanced, there exists two positive real numbers θ and p, and an integer k such that for all $n \geqslant k$ $a_n = \lfloor np + \theta \rfloor - \lfloor (n-1)p + \theta \rfloor$,*

Note that the sequence $a_n = \lfloor np + \theta \rfloor - \lfloor (n-1)p + \theta \rfloor$, is periodic in θ with period 1 and $\forall k$,

$$\lim_{d \to \infty} \frac{\sum_{i=k}^{k+d} a_i}{d} = p.$$

This justifies the fact that p is called the rate of a. Also note that $a_n \in \{\lfloor p \rfloor, \lceil p \rceil\}$.

The following lemma gives the relation between a balanced admission sequence a and its corresponding counting sequence b.

Lemma 7.3. *Let a be any integer sequence with values in $\{0, 1\}$ and b be the associated counting sequence ($b_n = \nu(n) - \nu(n-1)$). Then, a is balanced with rate p if and only if b is balanced with rate $1/p$.*

Proof. First, let us prove that if a is balanced, then b is balanced. Note that b is composed of only two integers n and $n+1$, otherwise, a would contain two sub-sequences, $10^{n-1}1$ and $10^{n-1+j}1$, with $j \geqslant 2$ (where 0^x stands for the sequence made of x consecutive 0's). This contradicts the balance condition for a given in Theorem 7.2. Now, let us suppose that b is not balanced. This means that there exists two sub-sequences of b of the same length k that contain i (resp. $i + j$) times the integer $(n+1)$, with $j \geqslant 2$. For the sequence a, this yields a first sub-sequence of length $kn + i + 1$ with $k + 1$ times the integer 1 and a second sequence of length $kn + i + j - 1$ that contains $k - 1$ times the integer 1. Since $j \geqslant 2$, the length of the second sub-sequence is larger than the length of the first one. This contradicts the fact that a is balanced.

Now let us suppose that b is balanced. By definition of a balanced sequence, b has all its values in $\{n, n+1\}$, for some n. Let us assume that a is not balanced. Using Theorem 7.2, this

[2] $\lfloor x \rfloor$ denotes the largest integer smaller or equal to x and $\lceil x \rceil$ the smallest integer which is not smaller than x

means that a contains two sub-sequences of the same length l, with respectively k and $k+j$ times the integer 1 and $j \geqslant 2$. The sub-sequence with the smaller number 1's is extended on both sides so that it contains exactly $k+j$ ones, starting with a 1 and ending with a 1. The second sub-sequence is cut on both side so that it starts and ends with a 1. Their respective length are now $l_1 \geqslant l+2 \geqslant l_2+2$. For b this yields two sub-sequences of the same length:$k+j-1$. They contain b_1 (resp. b_2) times the integer $n+1$. This means that for the sub-sequences in a, their lengths are respectively: $l_1 = 1+(k+j-1)n+b_1$ and $l_2 = 1+(k+j-1)n+b_1$. The fact that $l_1 \geqslant l+2 \geqslant l_2+2$ implies that $b_1 \geqslant b_1+2$. This contradicts the fact that b is balanced.

At this point we have established that a is balanced if and only if b is also balanced. As for the rates, assume that a has rate p, Then if we consider a sub-sequence in a with k times 1, its length is $l = k/p + o(k)$. For b, the associated interval is of length $k+o(k)$ and sums to $l-k+o(k)$. The rate is $l/k = 1/p$. □

7.2 The time slot approach

Now, we are ready to consider the problem to find the optimal admission policy with a given rate. We denote by $\mathbf{E}_{\sigma,\delta}(.)$ the expectation w.r.t. the inter-arrival times and the service times in all nodes in the system. Following the technical conditions given in [2], we have to make sure that the expected traveling times to node i under the admission sequence a, satisfy the properties given in Lemma 7.4. Let $h_i : \mathbb{R} \to \mathbb{R}, \quad 1 \leqslant i \leqslant Q$ be arbitrary non-decreasing convex functions. For simplicity we denote in the following $W_n^i(a_1, \cdots, a_n)$ instead of $h_i \circ W_n^i(a_1, \cdots, a_n)$.

In the next two lemmas, we use increasing and decreasing in the non-strict sense.

Lemma 7.4. *Assume that the inter-arrival times and the service times are stationary sequences, independent of each other. The following properties are true:*
i- $\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n)$ *is increasing.*
ii- $\mathbf{E}_{\sigma,\delta} W_n^i(a_{m-n+1}, \cdots, a_m) \leqslant \mathbf{E}_{\sigma,\delta} W_m^i(a_1, \cdots, a_m)$, $n < m$.
iii- $\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n) = \mathbf{E}_{\sigma,\delta} W_m^i(0, \cdots, 0, a_1, \cdots, a_n)$, $n < m$.
iv- $\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n)$ *is multimodular.*

Proof. We prove the four properties in a convenient order.

i- Using the extended Lindley formula, the expected traveling time $\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n)$ is increasing in a_k, $k = 1, \cdots n$.

iii- Let us fix n and m with $n < m$. Since the inter-arrival times are stationary, we can couple the δ's such that $\forall 1 \leqslant j \leqslant n$, $\delta_j^{(n)} = \delta_{j+m-n}^{(m)} = \delta_{j+m-n}$. Under this coupling and because the system is initially empty, we have:

$$W_n^i(a_1, \cdots, a_n) = W_m^i(0, \cdots, 0, a_1, \cdots, a_n),$$

and,

$$\mathbf{E}_\sigma W_n^i(a_1, \cdots, a_n) = \mathbf{E}_\sigma W_m^i(0, \cdots, 0, a_1, \cdots, a_n).$$

Therefore, if $n < m$,

$$\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n) = \mathbf{E}_{\sigma,\delta} W_m^i(0, \cdots, 0, a_1, \cdots, a_n).$$

ii- This is a direct consequence of *i* and *iii*.

iv- Theorem 5.4 shows that $\mathbf{E}_\sigma W_n^i(a_1, \cdots, a_n)$ is multimodular, thus $\mathbf{E}_{\sigma,\delta} W_n^i(a_1, \cdots, a_n)$ is multimodular as well. □

Now, we choose the cost function $g^i(a)$ of the control sequence $a = (a_1, a_2, \cdots)$ to be the Cezaro sum of the expectation of an increasing convex function of traveling time to a given node i, of a potential customer admitted to the queue at time T_n. We have:

$$g^i(a) \overset{\text{def}}{=} \overline{\lim_{N \to \infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\sigma,\delta} W_n^i(a_1, ..., a_n).$$

Note that $g^i(a)$ could be called the average traveling time at epochs $\{T_n\}$. In the case the sequence $\{T_n\}$ is a Poisson process, $g^i(a)$ is the time-average of the traveling times.

To find which admission policy minimizes the cost $g^i(.)$, we recall the notation for the balanced sequence with phase θ and rate p,

$$a_k^p(\theta) = \lfloor kp + \theta \rfloor - \lfloor (k-1)p + \theta \rfloor. \tag{36}$$

We can restate the general theorems given in [2] in the particular case studied here.

Theorem 7.5. *For each sequence a such that*

$$\varliminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_n \geqslant p,$$

it holds that,

$$g^i(a) \geqslant g^i(a^p(\theta)),$$

for any $\theta \in [0,1]$.

Proof. First note that the balanced sequence $a_k^p(\theta)$ has an asymptotic rate equal to p. Then, Theorem 3.5 in [2] can be applied, since all necessary assumptions on W are satisfied. □

7.3 The counting approach, the bounded case

In this section, we use the counting sequence b rather than the admission sequence a. Therefore, we rather focus on the function $h \circ \mathcal{W}(i)_n(b_1, \cdots, b_n)$, which is denoted $\mathcal{W}_n^i(b_1, \cdots, b_n)$ for the sake of notation simplicity.

Lemma 7.6. *Assume that the service time vectors and the inter-arrival times are stationary sequences, independent of each other. Assume that the σ_k^q (resp. δ_k), for each $k \in \mathbb{N}$ and $q \in \mathcal{Q}$ is bounded from above (resp. from below) by R (resp. by D).*
i- $\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n)$ is decreasing.
ii- $\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n) \geqslant \mathbf{E}_{\delta,\sigma}\mathcal{W}_m^i(b_{n-m+1}, \cdots, b_n)$, $n > m$.
iii- $\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n) = \mathbf{E}_{\delta,\sigma}\mathcal{W}_m^i(\beta, \cdots, \beta, b_{n-m+1}, \cdots, b_n)$, $n < m$, where $\beta = \lceil QR/D \rceil$.
iv- $\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n)$ is multimodular.

Proof. The proof holds by using stationarity of the inter-arrival times and of the service time vectors.

i- Using the extended Lindley's Formula, it is clear that $\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n)$ is decreasing.

ii- Let us fix n and m with $n > m$. Since the service time vectors are stationary, we can couple the service times such that $\forall 1 \leqslant j \leqslant m$, $\sigma_{j+n-m}^{(n)} = \sigma_j^{(m)} = \sigma_j$. We also couple the δ's in the following way:

$$\forall 1 \leqslant j \leqslant \nu(m),\ \delta_{j+\nu(n-m)}^{(n)} = \delta_j^{(m)} = \delta_j.$$

Under this coupling and because the system is initially empty, we have:

$$\mathcal{W}_n^i(b_1, \cdots, b_n) \quad \geqslant \quad \mathcal{W}_m^i(b_{n-m+1}, \cdots, b_n).$$

Therefore, the inequality holds for the expected values as well,

$$\mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1, \cdots, b_n) \geqslant \mathbf{E}_{\delta,\sigma}\mathcal{W}_m^i(b_{n-m+1}, \cdots, b_n).$$

Note that properties i and ii do not use the fact that the service time vectors and the inter-arrival times are bounded. This assumption is used for property *iii*.

iii- First, we show that under the sequence $(\beta, b_2, \cdots, b_n)$, the net is empty at the time of the second arrival.

Let us assume for a moment that the second arrival is infinitely delayed. Since the system is originally empty, then after the first arrival, each transition in the system will fire exactly once (this is a well known result for Event Graphs, see for example [7]). At this time, the system is again empty and no transition can fire. An upper bound of this time is given by QR. If the next customer arrives in the system more than QR units of time later, then it will find the system in the original empty state.

Now, note that $\tau_1 \geqslant \beta D$ by definition of D and taking $\beta \geqslant QR/D$ makes τ_1 larger than QR.

Finally, using the same coupling as for *ii*, we get

$$\mathbf{E}_{\delta,\sigma}\mathcal{W}^i_n(b_1, \cdots, b_n) = \mathbf{E}_{\delta,\sigma}\mathcal{W}^i_{n+1}(\beta, b_2, \cdots, b_n).$$

An easy induction on n gives the result.

iv- Theorem 6.4 shows that $\mathbf{E}_{\delta}\mathcal{W}^i_n(b_1, \cdots, b_n)$ is multimodular, thus $\mathbf{E}_{\delta,\sigma}\mathcal{W}^i_n(b_1, \cdots, b_n)$ is multimodular as well. □

Similarly to the previous case, we choose the cost function of any control sequence $b = (b_1, b_2, \cdots)$ to be the Cezaro sum of the expected traveling times of all customers admitted to the queue, to a given node i:

$$\gamma^i(b) = \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma}\mathcal{W}^i_n(b_1, ..., b_n).$$

We recall that the balanced policy with phase θ and rate p is

$$b^r_k(\theta) \stackrel{\text{def}}{=} \lfloor kr + \theta \rfloor - \lfloor (k-1)r + \theta \rfloor. \tag{37}$$

We can restate the general theorems given in [2] in the particular case studied here.

Theorem 7.7. *Under the foregoing assumptions, for all sequence b such that*

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} b_n \leqslant r,$$

then $\gamma^i(b) \geqslant \gamma^i(b^r(\theta))$, for any $\theta \in [0, 1]$.

Proof. Lemma 7.3 shows that the counting sequence associated with $b^r_k(\theta)$ is a balanced sequence with rate r.

Then, the optimization theory developed in [2] can be applied since all necessary assumptions on $\mathbf{E}_{\sigma,\delta}\mathcal{W}$ are satisfied. □

Note that Lemma 7.3 says that the average workload and the average waiting time are both optimized by the same admission control sequence $a^p(\theta)$.

In the next section, we will prove that the same optimization result holds in the unbounded case. The proof uses strongly the result stated in Theorem 7.7.

7.4 The unbounded case

If the service times and the inter-arrival times are not bounded in the original stochastic event graph $\mathcal{G}$, then, we use a fixed quantity Z and we introduce a new system $\mathcal{G}^Z$ where all services times, σ^q_n are replaced by $\sigma^{Z,q}_n \stackrel{\text{def}}{=} \min(Z, \sigma^q_n)$. The inter-arrival times δ_k are also replaced by $\delta^Z_k \stackrel{\text{def}}{=} \max(1/Z, \delta_k)$. In the new system $\mathcal{G}^Z$, the service times are bounded from above by Z and the inter-arrival times are bounded from below by $1/Z$.

In $\mathcal{G}^Z$, $\mathcal{W}^{Z,i}_n$ is the traveling time of the n-th customer to the i-th node.

If Θ is a random variable with uniform distribution on $[0,1)$,

$$b_k^r(\Theta) \stackrel{\text{def}}{=} \lfloor kr + \Theta \rfloor - \lfloor (k-1)r + \Theta \rfloor$$

is called the randomized balanced policy with rate r.

Following the results in [2], the time average of $\mathbf{E}_{\delta,\sigma}\mathcal{W}^Z$, is minimized by the randomized balanced policy, (as well by the one with θ fixed at an arbitrary value). We have for all policy b with rate less than p,

$$\begin{aligned}
\gamma^{Z,i}(b) &= \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma}\mathcal{W}_n^{Z,i}(b_1, ..., b_n) \\
&\geqslant \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma,\Theta}\mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta))
\end{aligned}$$

To prove the optimality of the randomized balanced sequence for the original system $\mathcal{G}$ where the service times and the inter-arrival times are unbounded, we need to let Z go to infinity in the previous inequality. For that, we need several technical lemmas.

Lemma 7.8. *The random sequence $b^r(\Theta)$ with Θ uniform on $[0,1)$ is stationary and $\mathbf{P}(b_k^r(\Theta) = 1) = r$, $\forall k$.*

Proof. By definition, $b_k^r(\Theta) = \lfloor kr + \Theta \rfloor - \lfloor (k-1)r + \Theta \rfloor$. The fact that $\int_0^1 \lfloor x + \theta \rfloor d\theta = x$ for any x implies that

$$\forall k, \ \int_0^1 b_k^r(\theta)d\theta = r.$$

□

Lemma 7.9. *If $\{\sigma_n\}$ is a stationary sequence of vectors, then $\{\min(\sigma_n, Z)\}$ and $\{\max(\sigma_n, Z)\}$ are also stationary.*

Proof. The proof follows by definition of stationary sequences □

We define the variable $\beta_k \stackrel{\text{def}}{=} \sum_{i=1}^k b_i$. The inter-admission times satisfy

$$\tau_k = \sum_{i=\beta_k}^{\beta_{k+1}-1} \delta_i. \tag{38}$$

One sees that τ_k is a sum of a random number of random variables.

Lemma 7.10. *Assume that the process $\{b_k\}$ is stationary, and the process $\{\delta_i\}$ is stationary and independent of $\{b_k\}$. Then, the process $\{\tau_k\}$ is stationary.*

Proof. We compute the distribution of the joint process $\tau_2, \cdots, \tau_k$. It is determined by the probabilities:

$$\begin{aligned}
P(\delta, b) &\stackrel{\text{def}}{=} \mathbf{P}(b_2 = n_2, \cdots, b_k = n_k, \delta_{b_1+1} \leqslant d_1, \cdots, \delta_{b_1+n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}) \\
&= \sum_i \mathbf{P}(b_1 = i, b_2 = n_2, \cdots, b_k = n_k, \delta_{i+1} \leqslant d_1, \cdots, \delta_{i+n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}).
\end{aligned}$$

Now,

$$
\begin{aligned}
P(\delta, b) &= \sum_i \mathbf{P}(b_i = i, b_2 = n_2, \cdots, b_k = n_k)\mathbf{P}(\delta_{i+1} \leqslant d_1, \cdots, \delta_{i+n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}) \\
&= \sum_i \mathbf{P}(b_1 = i, b_2 = n_2, \cdots, b_k = n_k)\mathbf{P}(\delta_1 \leqslant d_1, \cdots, \delta_{n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}) \\
&= \sum_i \mathbf{P}(b_0 = i, b_1 = n_2, \cdots, b_{k-1} = n_k)\mathbf{P}(\delta_1 \leqslant d_1, \cdots, \delta_{n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}) \\
&= \mathbf{P}(b_1 = n_2, \cdots, b_{k-1} = n_k, \delta_1 \leqslant d_1, \cdots, \delta_{n_1+\cdots+n_k} \leqslant d_{n_1+\cdots+n_k}),
\end{aligned}
$$

where the first equality follows from the independence of b and δ, the second follows from the stationarity of δ and the third from the stationarity of b.

This last expression gives the distribution of the joint process $\tau_1, \cdots \tau_{k-1}$. Since all of this holds for all k, the process $\{\tau_k\}$ is stationary. □

Lemma 7.11. *Under the randomized balanced policy, if $\{\sigma_k\}$ and $\{\delta_k\}$ are stationary sequences, then $\{\sigma_k^Z\}$ and $\{\tau_k^Z\}$ are also stationary.*

Proof. The proof comes from a direct combination of the three previous lemmas. □

Lemma 7.12. *If $\{\sigma_k^i\}$ and $\{\delta_k\}$ are stationary sequences and if the system $\mathcal{G}^Z$ is empty originally, then for all n,*

$$
\mathcal{W}_n^{Z,i}(b^p(\Theta)) \leqslant_s \mathcal{W}_{n+1}^{Z,i}(b^p(\Theta)),
$$

where $\leqslant_s$ denotes the stochastic order.

Proof. If the service time and the inter-arrival sequences are stationary (as it is the case here), this result is well known (see for example [5]). We can set $\mathcal{W}_0^{Z,i}(b^p(\Theta)) = 0$ and we have $\mathcal{W}_0^{Z,q}(b^p(\Theta)) \leqslant_s \mathcal{W}_1^{Z,i}(b^p(\Theta))$. Now, the Lindley formula for (max,+) systems can be used in this case.

$$
\mathcal{W}_k^Z(b^p(\Theta)) = H(k-1) \otimes D(-\tau_{k-1}) \otimes \mathcal{W}_{k-1}^Z(b^p(\Theta)) \oplus B,
$$

and

$$
\mathcal{W'}_{k+1}^Z(b^p(\Theta)) = H'(k) \otimes D(-\tau'_k) \otimes \mathcal{W'}_k^Z(b^p(\Theta)) \oplus B,
$$

where the primes denote another sample path of the service and inter-admission times. By induction, we can assume that component-wise, $\mathcal{W}_k^Z(b^p(\Theta)) \geqslant_s \mathcal{W}_{k-1}^Z(b^p(\Theta))$. Now, stationarity of the service times in all queues and of inter-arrival times comes from Lemma 7.11. We can couple the service times and the inter-arrival times such that for all i, $\sigma_k^{i'} = \sigma_{k-1}^i$ and $\tau'_k = \tau_{k-1}$. Under this coupling, $H(k-1) = H'(k)$ and $D(-\tau'_k) = D(-\tau_{k-1})$. Hence

$$
\mathcal{W'}_{k+1}^Z(b^p(\Theta)) \geqslant \mathcal{W'}_k^Z(b^p(\Theta))
$$

component-wise, under this coupling. □

The optimality of the randomized balanced policy is established by the following theorem.

Theorem 7.13. *Consider the system $\mathcal{G}$ where the service time vectors and the inter-arrival times are stationary sequences, independent of each other. For each control sequence b such that*

$$
\overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} b_n \leqslant p,
$$

it holds that

$$
\gamma^i(b) \geqslant \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta)).
$$

Proof. Remark that the quantity $\mathcal{W}_k^{Z,i}(b)$ is increasing in Z and

$$\lim_{Z\to\infty} \mathcal{W}_k^{Z,i}(b) = \sup_Z \mathcal{W}_k^{Z,i}(b) = \mathcal{W}_k^i(b).$$

The proof follows from the series of inequalities,

$$\begin{aligned}
\gamma^i(b) &= \overline{\lim_{N\to\infty}} \sup_Z \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma} \mathcal{W}_n^{Z,i}(b_1, ..., b_n) \right) \\
&\geqslant \sup_Z \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma} \mathcal{W}_n^{Z,i}(b_1, ..., b_n) \\
&\geqslant \sup_Z \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta)).
\end{aligned}$$

By Lemma 7.12, we know that $\mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta))$ is increasing in n. Therefore, the Cezaro limit equals the supremum on all n. We continue the previous inequalities:

$$\begin{aligned}
\gamma^i(b) &\geqslant \sup_Z \sup_n \mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta)) \\
&= \sup_n \sup_Z \mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{Z,i}(b_1^p(\Theta), ..., b_n^p(\Theta)) \\
&= \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma,\Theta} \mathcal{W}_n^{i}(b_1^p(\Theta), ..., b_n^p(\Theta)).
\end{aligned}$$

□

Our last theorem shows that the (non-randomized) balanced policy is optimal for any initial phase θ.

Theorem 7.14. *Under the assumptions of Theorem 7.13, for each control sequence b such that*

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} b_n \leqslant p,$$

it holds that

$$\gamma^i(b) \geqslant \gamma^i(b_p(\theta)),$$

for all $0 \leqslant \theta \leqslant 1$.

Proof. Define

$$f_m(\theta, p) \stackrel{\text{def}}{=} \mathbf{E}_{\delta,\sigma} \mathcal{W}_m^i(b_1^p(\theta), ..., b_m^p(\theta)).$$

Note that f_m is periodic (in θ) with period 1. Define

$$f'_m(\theta, p) \stackrel{\text{def}}{=} f_m(a_{-m+1}^p(\theta), ..., a_0^p(\theta)).$$

Then we have

$$f'_m(\theta', p) = f_m(\theta, p) \quad \text{where} \quad \theta' = \theta - mp, \tag{39}$$

Again f'_m is periodic w.r.t. θ, with period 1, and is increasing in m because of property *ii* in Lemma 7.4, so that the following limit exists,

$$f'_\infty(\theta, p) \stackrel{\text{def}}{=} \lim_{m\to\infty} f'_m(\theta, p).$$

Hence,

$$E_\Theta f'_\infty(\Theta,p) = \lim_{m\to\infty} E_\Theta f'_m(\Theta,p) = \lim_{m\to\infty} E_\Theta f_m(\Theta,p). \tag{40}$$

Now, Theorem 3.2 in [2] shows that under properties *ii* and *iv* of Lemma 7.4, the following holds,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbf{E}_{\delta,\sigma}\mathcal{W}_n^i(b_1^p(\theta),...,b_n^p(\theta)) = E_\Theta f'_\infty(\Theta,p).$$

Combining this with Theorem 7.13 and Equation 40 finishes the proof. □

8 Conclusion

We investigated in this paper the admission control problem in a general setting. We were motivated by the pioneering work of Hajek [10], who solved the admission control problem into a single queue in a Markovian framework. We study the admission into a general network having a (max,+) linear structure, under a general probabilistic framework, only stationarity and ergodicity assumptions are used instead of a Markovian setting.

Structural properties of the waiting time of any (max,+) linear system are established. In particular, we show that the waiting time seen as a function of the admission control sequence is multimodular. We described the optimal admission policy among all admission sequence with a given long range acceptance range.

In a forthcoming paper [1], we use again the framework established in [2] to another class of optimal control problems: the routing in systems composed of several (max,+) linear systems.

It would be important for applications if the results of this paper could be extended to queuing networks which have

1. more servers at each queue,
2. random routing of jobs,
3. different classes of jobs,
4. performance measures which are multidimensional convex functions of all the workloads (or traveling times) in the network (to analyze correlations, for instance).

Acknowledgment

The authors would like to thank François Baccelli for his help in Section 7.4 concerning stationarity issues, as well as an anonymous referee whose comments and insight helped improve several proofs of the paper.

References

[1] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. Technical Report 3180, INRIA, June 1997.

[2] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. Technical Report 3181, INRIA, June 1997.

[3] F. Baccelli. Ergodic theory of stochastic Petri networks. *Annals of Probability*, 20(1):375–396, 1992.

[4] F. Baccelli, G. Gohen, G.J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity*. Wiley, 1992.

[5] F. Baccelli and Z. Liu. On a class of stochastic recursive equations arising in queueing theory. *Annals of Probability*, 21(1):350–374, 1992.

[6] F. Baccelli and V. Schmidt. Taylor series expansions for poisson driven (max,+) linear systems. *The annals of Applied Probability*, 6(1):138–185, 1996.

[7] M. Canales and B. Gaujal. Marking optimization and parallelism in marked graphs. Technical Report 2049, INRIA, 1993.

[8] J. Desel and J. Esparza. *Free Choice Petri Nets.* Cambridge Tracts in Theorical Computer Science, 1995.

[9] P. Glasserman and D.D. Yao. Monotone optimal control of permutable gsmps. *Math. of Operation Research*, 19:449–476, 1994.

[10] B. Hajek. Extremal splittings of point processes. *Mathematics of Operation Research*, 10(4):543–556, 1985.

[11] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, 1979.

[12] M. Morse and G. A. Hedlund. Symbolic dynamics II- Sturmian trajectories. *American Journal Math.*, 62:287–306, 1940.

[13] M. L. Puterman. *Markov decision processes.* Wiley Series in Probabilities and Mathematical Statistics. Wiley, 1994.

[14] L. I. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley-Interscience Series in Probability and Statistics. John Wiley & Sons, 1998.

[15] S. Stidham. Optimal control of admission to a queueing system. *IEEE Transaction on Automatic Control*, AC-30(8):705–713, 1985.