Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy

TU/e technische universiteit eindhoven

SPOR-Report 2004-14

Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy

I.J.B.F. Adan
G. Weiss

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, October 2004
The Netherlands

/department of mathematics and computing science

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, Statistics and Operations research
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Main Building 9.10
Telephone: + 31 40 247 3130
E-mail: wscosor@win.tue.nl
Internet: http://www.win.tue.nl/math/bs/cosor.html

ISSN 1567-5211

Analysis of a Simple Markovian Re-Entrant Line with Infinite Supply of Work under the LBFS Policy

Ivo Adan *
Department of Mathematics and Computing Science
Eindhoven University of Technology
HG 9.09, P.O. Box 513,
5600 MB Eindhoven, the Netherlands
iadan@win.tue.nl

Gideon Weiss *†
Department of Statistics
The University of Haifa
Mount Carmel 31905, Israel.
gweiss@stat.haifa.ac.il

October 2004

Abstract

We consider a two machine 3 step re-entrant line, with an infinite supply of work. The service discipline is last buffer first served. Processing times are independent exponentially distributed. We analyze this system, obtaining steady state behavior and sample path properties.

Keywords: Queueing, manufacturing, priority scheduling, Markovian multiclass queueing networks, last buffer first served discipline, infinite virtual buffers, steady state distributions, sample path properties, GI/M/1 queue.

1 Introduction

We consider a production system with two machines, and a 3 step production process, where each part is processed first by machine one for the first step, then by machine two for the second step, and finally again by machine one for the third step, before leaving the system. The processing times for each of the 3 steps are independent sequences of independent identically distributed random variables, with means m_i and rates $\mu_i = 1/m_i$, $i = 1, 2, 3$. This system is the simplest example of a re-entrant line (as defined by Kumar [6]), which in turn is a special case of a multi-class queueing network (as described by Harrison [4]). This particular system has previously been studied in [10, 3, 1].

*Research supported in part by Network of Excellence Euro-NGI
†Research supported in part by Israel Science Foundation Grant 249/02

It is known that if parts arrive at this system in a renewal stream, at rate α, then under the condition $\rho_1 = \alpha(m_1 + m_3) < 1, \rho_2 = \alpha m_2 < 1$ the queues of parts waiting for each step (buffer levels) are stable, and in fact the system is positive Harris recurrent, for any work conserving policy (Dai and Weiss [3]). It is also known that any re-entrant line with $\rho_i = \alpha \sum_{k \in C_i} m_i < 1$, $i = 1, \ldots, I$ (where the steps are $k = 1, \ldots, K$, and steps $k \in C_i$ are performed at machine i) has stable queues, and is positive Harris recurrent, under the LBFS (Last Buffer First Served) policy (Kumar and Kumar [7] and Dai and Weiss [3]).

If however the arrival rate α is high enough to equal the bottleneck processing rate, i.e., $\max\{\alpha(m_1 + m_3), \alpha m_2\} = 1$, then the system is weakly stable but not stable: The departure rate from the queues is equal to α, but as time increases, the queue length at some of the buffers will converge weakly to infinity. Thus such a system cannot work at a rate $\max\{\rho_1, \rho_2\} = 1$, without accumulating unbounded queues.

In this note we consider a different situation, which is typical of manufacturing systems. We assume that there is an infinite supply of work available, so that there are always parts ready for processing step 1. In that case machine 1 will always be busy. We investigate the stability of this system under LBFS policy: Last buffer first served means here simply that machine 1 gives priority to parts in buffer 3 over parts in buffer 1. We assume that this priority is preemptive — whenever a part arrives in buffer 3, machine 1 will preempt the part in buffer 1 and start processing the part in buffer 3, and it will resume work on the part in buffer 1 only when buffer 3 is empty.

In this paper we consider only the case that $m_2 < m_1 + m_3$. In this case, if machine 1 works all the time and the system is weakly stable, then machine 2 will have a traffic intensity of $\frac{m_2}{m_1+m_3} < 1$.

For the sake of completeness we say a few words here about the case of $m_2 \geq m_1 + m_3$. If $m_2 \geq m_1 + m_3$ and if machine 1 works all the time, then the arrival rate into machine 2 will be $\geq \frac{1}{m_1+m_3} \geq \frac{1}{m_2}$. Hence it does not make sense for machine 1 to work all the time, and a sensible policy is to idle machine 1 if buffer 3 is empty, and buffer 2 is above a certain threshold level. This will always result in a departure rate $< \frac{1}{m_2}$ from the system, although it can be made arbitrarily close to $\frac{1}{m_2}$ if the threshold is raised. On the other hand, when buffer 2 is below a certain level, and buffer 3 is not too full, it may be sensible to stop serving buffer 3 and serve buffer 1 to replenish buffer 2 to avoid future starvation of machine 2. Thus one could have a switching curve, $S(n_3)$ where $n_3 = 0, 1, \ldots$ is the level of buffer 3, and $S(n_3)$ is a decreasing function of n_3, so that the policy is to serve buffer 1 if buffer 2 is below $S(n_3)$, and serve buffer 3 or idle if buffer 2 is at or above $S(n_3)$. It would be interesting to analyze such a system and choose a suitable switching curve — we leave this for future research.

We focus on the case that all the processing times are exponentially distributed. In this case the system can be described by a continuous time discrete state space Markov chain. In a recent paper Weiss [12] has shown that this chain is positive recurrent under the condition that $m_1 + m_3 > m_2$. In the present paper we provide a more detailed analysis: We obtain the steady state distribution of the chain, derive sample path properties, and analyze how the values of the parameters influence system performance.

Our results have some practical applications in job-shop scheduling heuristics; for further explanations and numerical experiments, see [11, 8].

2 Definition of the system and summary of results

Our re-entrant line manufacturing system is described schematically in Fig 1. Processing times

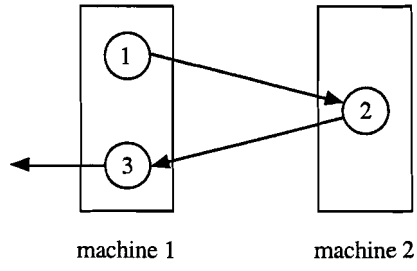

Figure 1: A 2 machine 3 step system, with virtual infinite buffer

at step i are i.i.d exponentially distributed with mean m_i, rate $\mu_i = 1/m_i$, for $i = 1, 2, 3$ and the three sequences are independent.

There are always parts available for processing of step 1. When parts finish processing step 1 by machine 1, they queue in buffer 2 where they remain until they are processed by machine 2 for step 2, and then they move to buffer 3, where they remain until they are processed by machine 1 for step 3, at which time they leave the system. Each buffer is processed in FIFO order. Processing is non-idling, that is, a machine will always process a part when there is work. We assume that machine 1 gives preemptive priority to buffer 3: Whenever there are parts in buffer 3, machine 1 will work on the first of them. When buffer 3 empties, machine 1 will immediately resume processing of a part in step 1. This is possible by the assumption that there is an infinite supply of work. We can think of it as if buffer 1 has an infinite queue of parts waiting for step 1. We call such a buffer *a virtual infinite queue.* The queue is virtual, because in practice buffer 1 need not contain many parts, but it needs to be monitored so it will never be empty. If during the processing of step 1 a part arrives from buffer 2 into buffer 3, machine 1 will preempt its work at buffer 1, and immediately start processing buffer 3.

Since the processing times are exponential we can describe this system as a discrete state continuous time Markov jump process, with the state given by the number of parts in buffers 2,3, denoted n_2, n_3. The state of the system at time t is $Q(t) = (Q_2(t), Q_3(t)) = (n_2, n_3), t \geq 0$. The transition rates of $Q(t)$ are presented in Fig. 2. They are:

$$\begin{aligned} (n_2, n_3) &\rightarrow (n_2 - 1, n_3 + 1) \text{ at rate } \mu_2, \quad n_2 > 0, \\ (n_2, n_3) &\rightarrow (n_2, n_3 - 1) \text{ at rate } \mu_3, \quad n_3 > 0, \\ (n_2, 0) &\rightarrow (n_2 + 1, 0) \text{ at rate } \mu_1, \quad n_2 \geq 0. \end{aligned} \tag{2.1}$$

In the remaining three sections of this paper we investigate this system. In Section 3 we obtain the steady state distribution of the system. It is quite close to a product form

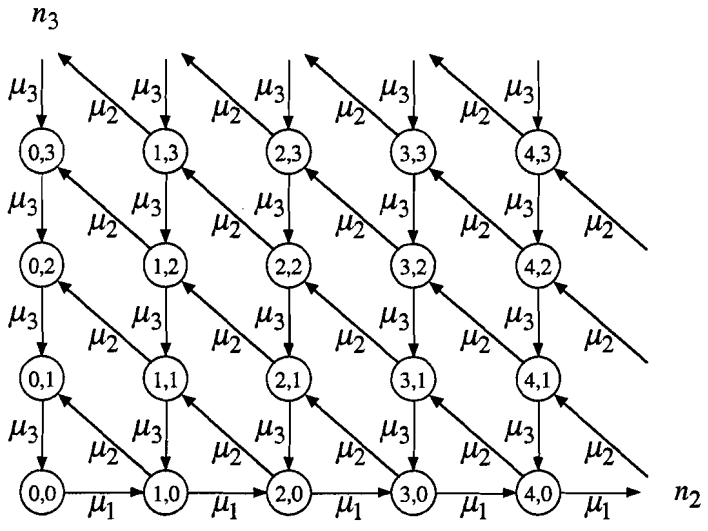

Figure 2: Transition rates for the Markovian states of the re-entrant system

distribution, in fact for $n_2, n_3 > 0$:

$$\lim_{t\to\infty} \mathbb{P}(Q_2(t) = n_2, Q_3(t) = n_3 | Q_2(t) > 0, Q_3(t) > 0) = (1-\alpha_2)\alpha_2^{n_2-1}(1-\alpha_3)\alpha_3^{n_3-1}, \quad n_2, n_3 = 1, 2, \ldots. \tag{2.2}$$

The marginal steady state distributions of both Q_2 and Q_3 are that of a GI/M/1 system, which is intriguing, since the arrival streams into Q_2 and into Q_3 are not independent.

In Section 4 we derive sample path properties of the process: Machine 1 will undergo cycles of work on buffer 1, which we call push periods, and work on buffer 3, which we call pull periods. Machine 2 will alternate between busy and idle periods. We derive properties of the length of these periods, and of the state of the system at the start and the end of these periods.

In Section 5 we analyze the dependence of system performance on parameter values. For m_2 we show that as m_2 increases from 0 to $m_1 + m_3$ the steady state contents of both buffers increase, and we derive the limiting steady state distributions as $m_2 \searrow 0$ and as $m_2 \nearrow m_1+m_3$. For m_1, m_3 we consider $m_1 + m_3 = a_1$ constant and we show that as m_3/m_1 increases, again the steady state contents of both buffers increase, and we derive the limiting steady state distributions as $m_3 \searrow 0$ and as $m_3 \nearrow a_1$. Of particular interest is the last case: When m_2 is fixed, and $a_1 = m_1 + m_3 > m_2$ is fixed, and we let $m_1 \searrow 0$ and $m_3 \nearrow a_1$, the queue lengths explode. If we then scale space and time by m_1 we obtain a *a stable but random fluid limit.*

3 Steady state distribution

Our main result in this section is the solution of the balance equations to obtain the steady state distribution of the network, when $m_1 + m_3 > m_2$.

Theorem 3.1 *The steady state distribution of the Markov jump process $Q(t)$ for the case that $m_1 + m_3 > m_2$ is:*

$$\lim_{t\to\infty} \mathbb{P}(Q(t) = (n_2, n_3)) = \begin{cases} \frac{m_1}{m_1+m_3}(1-\alpha_2)\alpha_2^{n_2}\alpha_3^{n_3}, & n_2 > 0 \text{ and } n_3 \geq 0 \text{ or } (n_2, n_3) = (0,0), \\ \frac{m_3}{m_1+m_3}(1-\alpha_2)\alpha_3^{n_3-1}, & n_2 = 0 \text{ and } n_3 > 0, \end{cases} \tag{3.1}$$

where:

$$\alpha_3 = \frac{\mu_1+\mu_2+\mu_3-\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}}{2\mu_3}, \tag{3.2}$$

$$\alpha_2 = \frac{\mu_1}{\mu_2}\frac{-\mu_1-\mu_2+\mu_3+\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}}{2\mu_3}. \tag{3.3}$$

Proof. The balance equations for the steady state probabilities are:

$$(\mu_2+\mu_3)P(n_2,n_3) = \mu_3 P(n_2,n_3+1)+\mu_2 P(n_2+1,n_3-1), \qquad n_2,n_3>0, \tag{3.4}$$

$$(\mu_1+\mu_2)P(n_2,0) = \mu_3 P(n_2,1)+\mu_1 P(n_2-1,0), \qquad n_2>0, \tag{3.5}$$

$$\mu_3 P(0,n_3) = \mu_3 P(0,n_3+1)+\mu_2 P(1,n_3-1), \qquad n_3>0, \tag{3.6}$$

$$\mu_1 P(0,0) = \mu_3 P(0,1). \tag{3.7}$$

We first solve the equations (3.4, 3.5), for all $n_2 > 0, n_3 \geq 0$. We use as trial solution: $\alpha_2^{n_2}\alpha_3^{n_3}$. Substituting in (3.4) and cancelling $\alpha_2^{n_2}\alpha_3^{n_3-1}$, and substituting in (3.5) and cancelling $\alpha_2^{n_2-1}$, we get two equations for α_2, α_3:

$$(\mu_2+\mu_3)\alpha_3 = \mu_3\alpha_3^2+\mu_2\alpha_2, \tag{3.8}$$

$$(\mu_1+\mu_2)\alpha_2 = \mu_3\alpha_2\alpha_3+\mu_1. \tag{3.9}$$

From (3.9) we get α_2 in terms of α_3:

$$\alpha_2 = \frac{\mu_1}{\mu_1+\mu_2-\mu_3\alpha_3}. \tag{3.10}$$

Note from (3.10) that:

$$0<\alpha_2<1 \quad \Leftrightarrow \alpha_3 < \frac{\mu_2}{\mu_3}. \tag{3.11}$$

Substituting (3.10) in (3.8), and multiplying by $\mu_1+\mu_2-\mu_3\alpha_3$ we get a cubic equation for α_3, one of whose roots is $\alpha_3 = \frac{\mu_2}{\mu_3}$:

$$f(\alpha_3) = (\mu_1+\mu_2-\mu_3\alpha_3)((\mu_2+\mu_3)\alpha_3-\mu_3\alpha_3^2)-\mu_1\mu_2 = \tag{3.12}$$

$$= \mu_3^2\alpha_3^3-\mu_3(\mu_1+2\mu_2+\mu_3)\alpha_3^2+(\mu_1+\mu_2)(\mu_2+\mu_3)\alpha_3-\mu_1\mu_2 = \tag{3.13}$$

$$= (\mu_3\alpha_3-\mu_2)(\mu_3\alpha_3^2-(\mu_1+\mu_2+\mu_3)\alpha_3+\mu_1) = 0. \tag{3.14}$$

If we take the derivative of $f(\alpha_3)$ and evaluate it at the root $\alpha_3 = \frac{\mu_2}{\mu_3}$ we get:

$$f'(\alpha_3)\big|_{\alpha_3=\frac{\mu_2}{\mu_3}} = 3\mu_3^2\alpha_3^2-2\mu_3(\mu_1+2\mu_2+\mu_3)\alpha_3+(\mu_1+\mu_2)(\mu_2+\mu_3)\big|_{\alpha_3=\frac{\mu_2}{\mu_3}} = \tag{3.15}$$

$$= 3\mu_2^2-2\mu_2(\mu_1+2\mu_2+\mu_3)+(\mu_1+\mu_2)(\mu_2+\mu_3) =$$

$$= \mu_1\mu_3-\mu_1\mu_2-\mu_2\mu_3 = \mu_1\mu_2\mu_3(m_2-m_1-m_3), \tag{3.16}$$

so that

$$f'(\alpha_3)\big|_{\alpha_3=\frac{\mu_2}{\mu_3}} < 0 \Leftrightarrow m_1 + m_3 > m_2. \tag{3.17}$$

Hence for the case that $m_1 + m_3 > m_2$ the remaining two roots of the cubic equation are both of them real, one of them smaller and one of them larger than the root at $\frac{\mu_2}{\mu_3}$. Since we are only interested in $0 < \alpha_2 < 1$ we will choose the smaller of these roots,

$$\alpha_3 = \frac{\mu_1 + \mu_2 + \mu_3 - \sqrt{(\mu_1 + \mu_2 + \mu_3)^2 - 4\mu_1\mu_3}}{2\mu_3}. \tag{3.18}$$

By $f(0) = -\mu_1\mu_2 < 0$, we have $\alpha_3 > 0$. It is also easily seen from (3.18) that $\alpha_3 < 1$.

Substituting (3.18) in (3.10) we get, after some manipulations:

$$\alpha_2 = \frac{\mu_1}{\mu_2} \frac{-\mu_1 - \mu_2 + \mu_3 + \sqrt{(\mu_1 + \mu_2 + \mu_3)^2 - 4\mu_1\mu_3}}{2\mu_3} = \frac{\mu_1}{\mu_2}(1 - \alpha_3). \tag{3.19}$$

We now need to consider the balance equations (3.6), at $n_2 = 0$, $n_3 > 0$. Note that none of $P(0, n_3)$, $n_3 > 0$, appear in the equations (3.4, 3.5) for $P(n_2, n_3)$, $n_2 > 0$, $n_3 \geq 0$. To solve (3.6) for $P(0, n_3)$, $n_3 > 0$, we now substitute the solution: $P(1, n_3 - 1) = \alpha_2\alpha_3^{n_3-1}$, $n_3 > 0$:

$$\mu_3 P(0, n_3) = \mu_3 P(0, n_3 + 1) + \mu_2\alpha_2\alpha_3^{n_3-1}, \qquad n_3 > 0. \tag{3.20}$$

This is a simple difference equation, for which we try the solution $c\alpha_3^{n_3}$. Substituting this trial solution, and cancelling $\alpha_3^{n_3-1}$ we get an equation for the value of c:

$$\mu_3 c\alpha_3 = \mu_3 c\alpha_3^2 + \mu_2\alpha_2, \tag{3.21}$$

from which (by the use of (3.19)):

$$c = \frac{\mu_2\alpha_2}{\mu_3\alpha_3(1 - \alpha_3)} = \frac{\mu_1}{\mu_3\alpha_3}. \tag{3.22}$$

Note that $c\alpha_3^{n_3}$ is a convergent solution. The general solution to the difference equation (3.20) is obtained by adding the general solution of the homogeneous equations. However, this is simply a constant, and so $c\alpha_3^{n_3}$ is the only convergent solution.

The remaining $P(0, 0)$ appears in equation (3.5) for $P(1, 0)$, and in equation (3.7). From the former we get:

$$P(0, 0) = (1 + \frac{\mu_2}{\mu_1})P(1, 0) - \frac{\mu_3}{\mu_1}P(1, 1), \tag{3.23}$$

from the latter equation we get:

$$P(0, 0) = \frac{\mu_3}{\mu_1}P(0, 1). \tag{3.24}$$

These two expressions for $P(0, 0)$ can be seen to agree, in view of (3.9). This is as expected by a well known general result, since one of the equations in the full set of balance equations is redundant.

We have derived a complete non-negative and convergent solution for the balance equations:

$$\begin{aligned} P(0,0) &= C, \\ P(0,n_3) &= C\frac{\mu_1}{\mu_3}\alpha_3^{n_3-1}, \qquad n_3 > 0 \\ P(n_2,n_3) &= C\alpha_2^{n_2}\alpha_3^{n_3}, \qquad n_2 > 0, n_3 \geq 0, \end{aligned} \tag{3.25}$$

and it remains to calculate the value of C which will normalize the sum of the probabilities to 1.

Adding up we get:

$$C^{-1} = 1 + \frac{\mu_1}{\mu_3}\frac{1}{1-\alpha_3} + \frac{\alpha_2}{1-\alpha_2}\frac{1}{1-\alpha_3}, \tag{3.26}$$

from which we get, after using (3.19) and (3.14):

$$C = \frac{m_1}{m_1 + \frac{m_2}{1-\alpha_2} + \frac{m_3}{1-\alpha_3}} = \frac{m_1}{m_1+m_3}(1-\alpha_2). \tag{3.27}$$

The last expression for the normalizing constant can also be verified indirectly. Summing up (3.25) over n_2 for $n_3 = 0$ we get:

$$\lim_{t\to\infty} \mathbb{P}(Q_3(t) = 0) = C\frac{1}{1-\alpha_2}.$$

But the steady state probability that $Q_3(t) = 0$ is the long term fraction of time that buffer 3 is empty, and this is exactly the fraction of time that machine 1 is working on buffer 1, which for a stable system equals $\frac{m_1}{m_1+m_3}$.

This completes the proof. ■

Immediately from the steady state distribution we get the following facts:

Marginal Steady State Distributions

$$\lim_{t\to\infty} \mathbb{P}(Q_2(t) = n_2) = \begin{cases} 1 - \frac{m_2}{m_1+m_3} & n_2 = 0, \\ \frac{m_2}{m_1+m_3}(1-\alpha_2)\alpha_2^{n_2-1} & n_2 > 0, \end{cases} \tag{3.28}$$

$$\lim_{t\to\infty} \mathbb{P}(Q_3(t) = n_3) = \begin{cases} \frac{m_1}{m_1+m_3} & n_3 = 0, \\ \frac{m_3}{m_1+m_3}(1-\alpha_3)\alpha_3^{n_3-1} & n_3 > 0. \end{cases} \tag{3.29}$$

Here the steady state probability that $Q_3 = 0$ equals the fraction of time that machine 1 is working on buffer 1. The steady state probability that $Q_2 = 0$ equals the fraction of idle time of machine 2 which results from $m_1 + m_3 > m_2$. The tails of the marginal distributions are geometric.

We recall that this is exactly the steady state distribution of a GI/M/1 queue, where the probability that the queue is empty is 1 minus the traffic intensity, and the probability of the queue length when it is not empty is geometric with parameter α which is the solution of the equation $\alpha = E(e^{-(1-\alpha)A})$ where A is the inter-arrival time random variable, see [2]. This is intriguing, because the queues at buffers 2 and 3 are not GI/M/1 — in fact the input streams into the buffers are far from GI inputs.

Conditional Joint Distributions

Conditional on both $Q_2 > 0$ and $Q_3 > 0$, the joint steady state distribution of the queues in both stations has the product form:

$$\lim_{t\to\infty} \mathbb{P}(Q_2(t) = n_2, Q_3(t) = n_3 | Q_2(t) > 0, Q_3(t) > 0) = (1-\alpha_2)\alpha_2^{n_2-1}(1-\alpha_3)\alpha_3^{n_3-1}, \quad n_2, n_3 = 1, 2, \ldots. \tag{3.30}$$

Moments of the Steady State Distributions

In steady state we have:

$$\begin{aligned}
\mathbb{E}(Q_2) &= \frac{m_2}{m_1+m_3}\frac{1}{1-\alpha_2}, \\
\mathbb{E}(Q_3) &= \frac{m_3}{m_1+m_3}\frac{1}{1-\alpha_3}, \\
\mathbb{V}(Q_2) &= \frac{m_2}{m_1+m_3}\frac{1+\alpha_2}{(1-\alpha_2)^2} - \left(\frac{m_2}{m_1+m_3}\right)^2 \frac{1}{(1-\alpha_2)^2}, \\
\mathbb{V}(Q_3) &= \frac{m_3}{m_1+m_3}\frac{1+\alpha_3}{(1-\alpha_3)^2} - \left(\frac{m_3}{m_1+m_3}\right)^2 \frac{1}{(1-\alpha_3)^2}, \\
\mathbb{C}\mathrm{ov}(Q_2, Q_3) &= \frac{m_2}{m_1+m_3}\left(\alpha_3 - \frac{m_3}{m_1+m_3}\right)\frac{1}{1-\alpha_2}\frac{1}{1-\alpha_3}.
\end{aligned} \tag{3.31}$$

4 Sample path properties

4.1 Push and pull periods

The process proceeds alternately through periods in which $n_3 = 0$ and machine 1 is processing buffer 1, and periods in which $n_3 > 0$ and machine 1 is processing buffer 3. We call the latter *pull periods*, in which the total number of parts in the system decreases, and the former *push periods*, in which the total number of parts in the system increases. Of course, in steady state

$$\frac{\mathbb{E}(PushPeriod)}{\mathbb{E}(PullPeriod)} = \frac{m_1}{m_3}.$$

Define as B_n the number of parts in buffer 2 at the beginning of the nth push period. Define as C_n the number of parts in buffer 2 at the end of the nth push period, just before a part moves from buffer 2 to buffer 3.

Clearly, $\{B_n, n = 0, 1, \ldots\}$ and $\{C_n, n = 1, 2, \ldots\}$, are Markov chains, and the sequence $B_0, C_0, B_1, C_1, \ldots$ is a Markov chain with alternating transition probabilities. Let $\mathbf{B} = (b_{ij})$ be the transition probabilites from B_n to C_n, and $\mathbf{C} = (c_{ij})$ be the transition probabilites from C_n to B_{n+1}.

Proposition 4.1

$$b_{ij} = \mathbb{P}(C_n = j | B_n = i) = \begin{cases} \frac{\mu_2}{\mu_1+\mu_2}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{j-i}, & j \geq i > 0, \\ \frac{\mu_2}{\mu_1+\mu_2}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{j-1}, & j > i = 0, \end{cases} \tag{4.1}$$

$$c_{ij} = \mathbb{P}(B_{n+1} = j | C_n = i) = \begin{cases} \binom{2(i-j)-1}{i-j} \frac{1}{2(i-j)-1} \frac{(\mu_2/\mu_3)^{i-j-1}}{(1+\mu_2/\mu_3)^{2(i-j)-1}}, & j = 1, 2, \ldots, i-1, \\ 1 - \sum_{k=1}^{i-1} \binom{2k-1}{k} \frac{1}{2k-1} \frac{(\mu_2/\mu_3)^{k-1}}{(1+\mu_2/\mu_3)^{2k-1}}, & j = 0. \end{cases} \tag{4.2}$$

Proof. If $B_n = 0$, then the push period will start with machine 2 idle until machine 1 completes the processing of one part out of buffer 1. The remainder of the push period consists of the processing time of one part by machine 2. During that time machine 1 will complete the processing of an additional L parts, where L is a geometric random variable,

$$\mathbb{P}(L = \ell) = \frac{\mu_2}{\mu_1 + \mu_2}\left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^{\ell}, \qquad l = 0, 1, \ldots.$$

This proves (4.1).

If a push period ends with i parts, then the pull period starts with $i-1$ parts in buffer 2 and 1 part in buffer 3 (the part that finished processing in buffer 2 and moved to buffer 3, thus starting the pull period). At this point buffer 3 acts like an M/M/1 queue, and the pull period corresponds to a busy period of that queue, truncated by the total number i. Let K be the number of customers served in a busy period of an M/M/1 queue, with arrival rate μ_2, and service rate μ_3. Then (see [9], Section 2.3.1 or [2] §II2.2, eq. 2.43):

$$\mathbb{P}(K = k) = \binom{2k-1}{k} \frac{1}{2k-1} \frac{(\mu_2/\mu_3)^{k-1}}{(1+\mu_2/\mu_3)^{2k-1}}, \qquad k = 1, 2, \ldots.$$

If a push period ended with i and the busy period serves k then at the end of the pull period there will be $j = (i-k)^+$ parts left in buffer 2. This proves (4.2). ■

Fig 3 illustrates the push and pull periods. Against the horizontal time axis we plot the level of buffer 2 on the positive side , and the level of buffer 3 on the negative side.

4.2 Steady state distribution at beginning and end of push period

We can easily obtain the steady state distributions of B_n and C_n from the steady state distribution of (Q_2, Q_3).

Proposition 4.2 *The steady state distribution of the level of buffer 2 at the beginning and the end of a push period are:*

$$\lim_{n\to\infty} \mathbb{P}(B_n = k) = \begin{cases} 1 - \frac{m_2}{m_3}\alpha_3, & k = 0, \\ \frac{m_2}{m_3}\alpha_3(1-\alpha_2)\alpha_2^{k-1}, & k = 1, 2, \ldots, \end{cases} \tag{4.3}$$

$$\lim_{n\to\infty} \mathbb{P}(C_n = k) = (1-\alpha_2)\alpha_2{}^{k-1}, \quad k = 1, 2, \ldots. \tag{4.4}$$

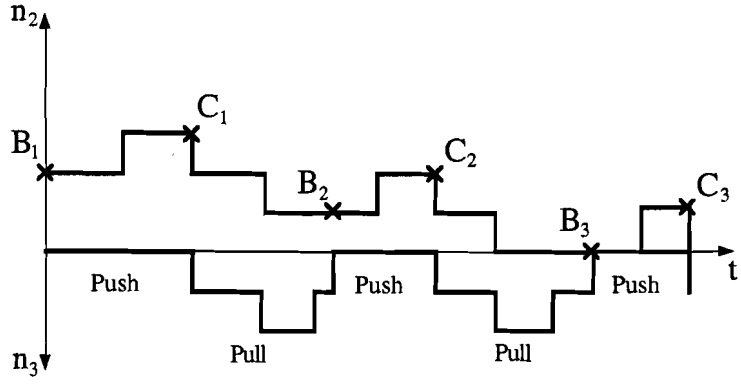

Figure 3: Push and Pull perioods

Proof. Consider first the end of a push period. The system always enters a pull period from state $(Q_2(t), Q_3(t)) = (k, 0), k > 0$ when machine 2 completes the processing of a part. The steady state rate at which this happens is $\mu_2 P(k, 0)$. Hence, the steady state rate at which the system enters pull periods is $\mu_2 \sum_{j=1}^{\infty} P(j, 0)$, and the steady state rate at which the system enters a pull period from a push period which ended at the level $C_n = k$ is $\mu_2 P(k, 0)$. Hence the long term fraction of push periods which end at level $C_n = k$ is $\mu_2 P(k, 0)/\mu_2 \sum_{j=1}^{\infty} P(j, 0)$. This long term fraction is the steady state probability of $C_n = k$.

Using the steady state probabilities (3.1),

$$P(k, 0) = \frac{m_1}{m_1 + m_3} (1 - \alpha_2)\alpha_2^k,$$

and (4.4) follows immediately.

Consider next the beginning of a push period. The system always enters a push period from state $(Q_2(t), Q_3(t)) = (k, 1), k \geq 0$ when machine 1 completes the processing of the last part in buffer 3. The same argument as for C_n shows that the steady state probability of $B_n = k$ is $\mu_3 P(k, 1)/\mu_3 \sum_{j=0}^{\infty} P(j, 1)$, where

$$P(k, 1) = \begin{cases} \frac{m_3}{m_1+m_3} (1 - \alpha_2), & k = 0, \\ \frac{m_1}{m_1+m_3} (1 - \alpha_2)\alpha_2^k \alpha_3, & k = 1, 2, \ldots. \end{cases}$$

Hence, we get that the steady state probability that B_n equals 0 is: $\frac{m_3(1-\alpha_2)}{m_1\alpha_2\alpha_3+m_3(1-\alpha_2)}$. It is now a matter of simple algebraic manipulations to show that this equals $1 - \frac{m_2}{m_3}\alpha_3$ and (4.3) follows immediately. ■

4.3 Steady state duration of push and pull periods

We denote by T_n^{Push} and T_n^{Pull} the duration of the nth Push period and the nth Pull period.

The durations of push periods are easy, we have trivially:

Proposition 4.3 *A push period which starts with $B_n > 0$ has duration $T_n^{Push} \sim \exp(\mu_2)$. A push period which starts with $B_n = 0$ has duration which is the sum of a buffer 1 and buffer 2 service times, $T_n^{Push} \sim \exp(\mu_1) * \exp(\mu_2)$ (convolution of two exponential distributions). The steady state distribution of a push period has probability density function:*

$$\lim_{n\to\infty} f(T_n^{Push} = t) = \frac{m_3 - m_1\alpha_3}{m_3(m_2 - m_1)} e^{-\frac{t}{m_2}} - \frac{m_3 - m_2\alpha_3}{m_3(m_2 - m_1)} e^{-\frac{t}{m_1}}. \tag{4.5}$$

We consider now the pull periods. A pull period which starts with $C_n = k$ consists of a busy period of the M/M/1 queue with input rate μ_2 and service rate μ_3, truncated by a total number of arrivals which does not exceed k. The distribution of this conditional duration of a pull period is not straightforward, and not very illuminating.

Instead we will calculate the steady state distribution of pull periods, i.e., the $n \to \infty$'s pull period. This steady state pull period will start from C_∞ which is $\sim$ geometric$(1 - \alpha_2)$ (see (4.4)). We need then to consider a pull period which starts with K parts, and is therefore truncated at K services, where K is a random variable, with geometric distribution with parameter $1 - \alpha_2$.

This K truncated busy period can be presented somewhat differently: Consider the M/M/1 queue in buffer 3, with input rate μ_2 and service rate μ_3, while the input source (buffer 2) is not yet empty. Then one of the following can happen: Either buffer 3 completes a service, with rate μ_3, and the queue in buffer 3 decreases, or an input occurs out of buffer 2 into buffer 3, which increases the queue in buffer 3 by 1. In the latter case, buffer 2 will decrease by 1. Since it started off with K parts, and was not empty before, there is a probability of α_2 that it is still not empty, and a probability $1-\alpha_2$ that the part that moved to buffer 3 was the last part, and buffer 2 is now empty. These two distinct events occur with rates $\mu_2\alpha_2$ and $\mu_2(1 - \alpha_2)$. We can then describe the state of the queue of buffer 3 by the number of parts in the buffer, plus an indicator of the state of buffer 2. This indicator is 1 if buffer 2 is still not empty and is 0 if buffer 2 is empty. The transition rates of this process are:

$$\begin{aligned}
(n_3, 1) &\to (n_3 - 1, 1) \text{ at rate } \mu_3, \quad n_3 > 0, \\
(n_3, 1) &\to (n_3 + 1, 1) \text{ at rate } \mu_2\alpha_2, \quad n_3 \geq 0, \\
(n_3, 1) &\to (n_3 + 1, 0) \text{ at rate } \mu_2(1 - \alpha_2), \quad n_3 \geq 0, \\
(n_3, 0) &\to (n_3 - 1, 0) \text{ at rate } \mu_3, \quad n_3 > 0.
\end{aligned} \tag{4.6}$$

For the continuous time Markov jump process defined by (4.6) we can define return times $T(n_3, i)$ as the time to return to states $(0, 0)$ or $(0, 1)$ when buffer 3 is empty, starting from state (n_3, i). In particular, the steady state pull period is the mixture random variable T given by

$$T = \begin{cases} T(1,1) & \text{w.p. } \alpha_2, \\ T(1,0) & \text{w.p. } 1 - \alpha_2. \end{cases}$$

We can now set up difference equations for the Laplace-Stieltjes transform (LST) of these return times, $\tilde{T}(n_3, i)$ (which is a function of a variable s but we drop s to simplify notation),

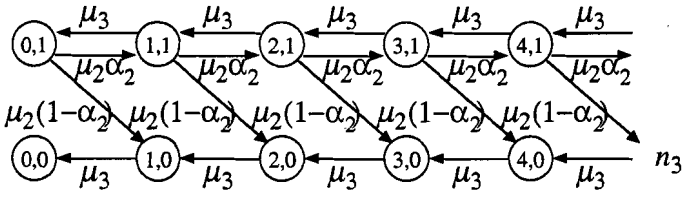

Figure 4: Transition rates for the marked buffer 3 process in a pull period

and thus obtain the LST of the steady state pull period duration. We can then invert the LST and obtain its distribution.

The system (4.6) stays in state $(k,1)$ for a duration $\sim \exp(\mu_2+\mu_3)$, and then transits to one of the states $(k-1,1)$, $(k+1,1)$, $(k+1,0)$. From each of those it will then have an independent remaining return time, hence:

$$\tilde{T}(k,1) = \frac{\mu_2+\mu_3}{\mu_2+\mu_3+s}\left[\frac{\mu_3}{\mu_2+\mu_3}\tilde{T}(k-1,1) + \frac{\mu_2\alpha_2}{\mu_2+\mu_3}\tilde{T}(k+1,1) + \frac{\mu_2(1-\alpha_2)}{\mu_2+\mu_3}\tilde{T}(k+1,0)\right], \quad k=1,2,3,\ldots.$$

Also, $T(k,0)$ is $\sim \text{Erlang}(k,\mu_3)$, and $T(0,1) = T(0,0) = 0$.

Hence the difference equations for the LSTs are:

$$(\mu_2+\mu_3+s)\tilde{T}(k,1) = \mu_3\tilde{T}(k-1,1) + \mu_2\alpha_2\tilde{T}(k+1,1) + \mu_2(1-\alpha_2)\tilde{T}(k+1,0), \quad k=1,2,3,\ldots, \tag{4.7}$$

with boundary terms:

$$\begin{aligned} \tilde{T}(0,1) &= 1, \\ \tilde{T}(k,0) &= \left(\frac{\mu_3}{\mu_3+s}\right)^k, \quad k \geq 0. \end{aligned} \tag{4.8}$$

Standard technique leads to the solution of these:

$$\tilde{T}(k,1) = (1-d)\lambda^k + d\left(\frac{\mu_3}{\mu_3+s}\right)^k, \quad k \geq 0, \tag{4.9}$$

where

$$\lambda = \frac{\mu_2+\mu_3+s-\sqrt{(\mu_2+\mu_3+s)^2-4\mu_2\mu_3\alpha_2}}{2\mu_2\alpha_2} \tag{4.10}$$

$$d = \frac{\mu_3(1-\alpha_2)}{\mu_3(1-\alpha_2)+s} \tag{4.11}$$

Note that $\alpha_3\lambda$ is exactly the LST of the busy period of an $M/M/1$ queue with arrival rate $\frac{\mu_2\alpha_2}{\alpha_3}$ and service rate $\mu_3\alpha_3$, see eq. 5.144 in [5]. We therefore have:

Proposition 4.4 *The LST of the stationary pull period is:*

$$\begin{aligned}\tilde{B}(s) &= \alpha_2\left((1-d)\lambda + d\left(\frac{\mu_3}{\mu_3+s}\right)\right) + (1-\alpha_2)\left(\frac{\mu_3}{\mu_3+s}\right) \\ &= \alpha_2(1-d)\lambda + d, \end{aligned} \tag{4.12}$$

and the density function of the steady state distribution of a pull period is:

$$\lim_{n\to\infty} f(T_n^{Pull} = t) = \alpha_2\left(l(t) - \int_{s=0}^{t} l(t-s)\mu_3(1-\alpha_2)e^{-\mu_3(1-\alpha_2)s}ds\right) + \mu_3(1-\alpha_2)e^{-\mu_3(1-\alpha_2)t}, \tag{4.13}$$

where $\alpha_3 l(t)$ is the density of the busy period distribution of an $M/M/1$ queue with arrival rate $\mu_2\alpha_2\alpha_3^{-1}$ and service rate $\mu_3\alpha_3$,

$$l(t) = \frac{\sqrt{\mu_3}}{t\sqrt{\mu_2\alpha_2}}\, e^{-(\mu_2\alpha_2\alpha_3^{-1}+\mu_3\alpha_3)t} I_1(2t\sqrt{\mu_2\mu_3\alpha_2}),$$

with $I_1(\cdot)$ denoting the modified Bessel function of the first kind of order one.

4.4 Machine 2 idle and busy period cycles

We denote by T_n^{Busy} and T_n^{Idle} the duration of the nth Busy period and the nth Idle period of machine 2. We first consider the idle periods. Along the same lines as in the proof of Proposition 4.2 it follows that the steady state distribution of the level of buffer 3 at the beginning of an idle period of machine 2 is $\sim$ geometric$(1-\alpha_3)$. Thus the time needed to empty buffer 3 is exponential with parameter $(1-\alpha_3)\mu_3$ and then, after a part has finished processing in buffer 1, the idle period of machine 2 ends. Hence we find:

Proposition 4.5 *The steady state distribution of an idle period of machine 2 is $\sim$ exp$((1-\alpha_3)\mu_3) * exp(\mu_1)$ with density function:*

$$\lim_{n\to\infty} f(T_n^{Idle} = t) = \frac{1-\alpha_3}{(1-\alpha_3)m_1 - m_3}\left[e^{-\frac{t}{m_1}} - e^{-\frac{(1-\alpha_3)t}{m_3}}\right]. \tag{4.14}$$

Now we consider the busy period of machine 2. It is, however, (too) hard to find its steady state distribution. Therefore we only derive the mean busy period duration, which immediately follows from the relation

$$\frac{m_2}{m_1+m_3} = \frac{\mathbb{E}T^{\text{Busy}}}{\mathbb{E}T^{\text{Busy}} + \mathbb{E}T^{\text{Idle}}}.$$

Proposition 4.6 *The steady state mean idle period and mean busy period of machine 2 are:*

$$\lim_{n\to\infty} \mathbb{E}T_n^{Idle} = \frac{m_3}{1-\alpha_3} + m_1, \tag{4.15}$$

$$\lim_{n\to\infty} \mathbb{E}T_n^{Busy} = \frac{m_2}{m_1+m_3-m_2}\left[\frac{m_3}{1-\alpha_3} + m_1\right]. \tag{4.16}$$

5 Monotonicity results

Having analyzed this simple 2 machine 3 buffer re-entrant line, we now investigate how system performance changes as a function of its parameters.

We consider first fixed m_1, m_3, and discuss how the performance of the system is affected by varying the value of m_2 in the range $0 < m_2 < m_1 + m_3$. We start with an informal discussion, based on insights from Section 4. Assume that m_2 is close to zero. Then typically, when a part will complete step 1, it will pass to buffer 2, but will only stay there for a short delay, and before another part arrives from buffer 1 the part will move to buffer 3. At that time machine 1 will switch to buffer 3 and will process the last step of this part. The part will then leave the system, and machine 1 will start to process step 1 of the next part. Hence, departures of parts from the system will be approximately the event times of a renewal process which is the convolution of the $\exp(\mu_1)$ and $\exp(\mu_3)$ processing times of steps 1 and 3, and buffer 2 will be empty almost all the time. On the contrary, if m_2 is close to $m_1 + m_3$ then the traffic intensity of buffer 2 will be close to 1, and therefore buffer 2 will have a large number of parts in it. Because buffer 2 is rarely empty, we will then have that buffer 3 behaves most of the time like an M/M/1 queue, with traffic intensity $m_3/(m_1 + m_3)$. In fact we show:

Proposition 5.1 *Consider fixed m_1, m_3.*

(i) As m_2 increases from 0 to $m_1 + m_3$, the steady state queue lengths in buffers 2 and 3 increase stochastically.

(ii)

$$\lim_{m_2 \searrow 0} \quad \lim_{t\to\infty} \mathbb{P}(Q_3(t) = n_3) = \begin{cases} \frac{m_1}{m_1+m_3}, & n_3 = 0 \\ \frac{m_3}{m_1+m_3}, & n_3 = 1, \end{cases} \tag{5.1}$$

$$\lim_{m_2 \searrow 0} \quad \lim_{t\to\infty} \mathbb{P}(Q_2(t) = 0) = 1, \tag{5.2}$$

$$\lim_{m_2 \nearrow m_1+m_3} \quad \lim_{t\to\infty} \mathbb{P}(Q_3(t) = n_3) = \frac{m_1}{m_1 + m_3} \left(\frac{m_3}{m_1 + m_3}\right)^{n_3}, \quad n_3 = 0, 1, 2, \ldots \tag{5.3}$$

$$\lim_{m_2 \nearrow m_1+m_3} \quad \lim_{t\to\infty} \mathbb{P}(Q_2(t) > n_2) = 1. \tag{5.4}$$

Proof. From (3.29) we have

$$\lim_{t\to\infty} \mathbb{P}(Q_3(t) > n_3) = \frac{m_3}{m_1 + m_3} \alpha_3^{n_3}, \quad n_3 = 0, 1, 2, \ldots.$$

Straight forward calculus and algebraic manipulations show that $\frac{d\alpha_3}{d\mu_2} < 0$ which proves that α_3 increases in $m_2 = 1/\mu_2$. It is also straightforward to show that α_3 converges to 0 when $m_2 \searrow 0$, and that it equals $\frac{m_3}{m_1+m_3}$ when $m_2 = m_1 + m_3$.

From (3.28) we have

$$\lim_{t\to\infty} \mathbb{P}(Q_2(t) > n_2) = \frac{m_2}{m_1 + m_3} \alpha_2^{n_2}, \quad n_2 = 0, 1, 2, \ldots.$$

To show that this is increasing in m_2 for every value of n_2, we show that α_2 is increasing with m_2. Straightforward though lengthy calculus and algebraic manipulations show that $\frac{d\alpha_2}{d\mu_2} < 0$ which proves that α_2 increases in $m_2 = 1/\mu_2$. Substituting $m_2 = 0$ we get $\alpha_2 = 0$, and substituting $m_2 = m_1 + m_3$ we get $\alpha_2 = 1$.

We provide the detailed proof in the Appendix. ■

We next consider fixed $a_1 = m_1 + m_3$ and fixed m_2, where $a_1 > m_2$, and we discuss how the system performance varies as we let m_3 increase from 0 to a_1 (while m_1 decreases from a_1 to 0). We again start with an informal discussion, based on insights from Section 4. Assume that m_3 is close to zero. In that case parts that leave buffer 2 experience very little delay (short pull periods), and machine 1 is processing step 1 of new parts most of the time (push period), hence buffer 3 is almost always empty, and buffer 2 behaves like an M/M/1 queue with input rate $\mu_1 \approx 1/a_1$ and service rate μ_2.

Assume at the other extreme that m_3 is close to a_1, so that m_1 is close to zero. Now we will have in particular that $m_3 > m_2$. Hence buffer 3 will behave like an unstable M/M/1 queue, while supplies of parts from buffer 2 last. At the same time input from buffer 1 to buffer 2, when buffer 3 is empty, will be very fast and so buffer 2 will experience highly bursty input. Typically, when both buffers are empty, buffer 2 will fill up with an initial burst (of duration $\sim \exp(\mu_2)$), and then there will be a geometric number of terminating busy periods of buffer 3 (where buffer 3 empties before buffer 2 runs out of parts), separated by similar bursts of parts moving from buffer 1 into 2. Eventually, processing in buffer 3 will fall behind the processing in buffer 2, and in a very long busy period buffer 2 will drain to zero. At this point buffer 3 will contain approximately a fraction $1 - m_2/m_3$ of all the parts which were produced out of buffer 1 since the last time that buffer 2 was empty. Buffer 3 will continue to process those remaining parts until it is empty, and during this period buffer 2 will be empty. When both buffers are empty the same cycle will repeat.

For all intermediate values of $0 < m_3 < m_1 + m_3$, buffer 2 has input rate $1/(m_1 + m_3)$ and service rate μ_2 but as m_3 increases the input is more bursty, which, as we shall see, increases the queue lengths at buffer 2. Naturally the queue length at buffer 3 increases with m_3.

Proposition 5.2 *Consider fixed $m_1 + m_3 = a_1$ and fixed m_2, where $m_2 < a_1$.*

(i) As m_3 increases from 0 to a_1 the steady state queue lengths in buffers 2 and 3 increase stochastically.

(ii)

$$\lim_{m_3 \searrow 0} \lim_{t\to\infty} \mathbb{P}(Q_3(t) = 0) = 1, \tag{5.5}$$

$$\lim_{m_3 \searrow 0} \lim_{t\to\infty} \mathbb{P}(Q_2(t) = n_2) = \frac{m_1}{m_1 + m_2} \left(\frac{m_2}{m_1 + m_2}\right)^{n_2}, \quad n_2 = 0, 1, 2, \ldots. \tag{5.6}$$

(iii) As $m_3 \nearrow a_1$ and $m_1 \searrow 0$ the fluid scaling limit of the queue length process

$$\bar{Q}(t) = \lim_{m_1 \searrow 0} m_1 Q(\frac{t}{m_1})$$

has regeneration instances where $(\bar{Q}_2(t), \bar{Q}_3(t)) = (0,0)$, *at which a vertical jump occurs in* $\bar{Q}_2$, *which is of size* $Y \sim \exp(\frac{1}{m_2} - \frac{1}{a_1})$. *This is followed by linear emptying of* $\bar{Q}_2$ *at rate* $\frac{1}{m_2}$, *and of* $\bar{Q}_2 + \bar{Q}_3$ *at rate* $\frac{1}{a_1}$. *If* T_n *is a regenration time, and* Y *is the jump size, then the fluid process following* T_n, *until* $T_{n+1} = T_n + a_1 Y$ *is given by:*

$$\bar{Q}_2(T_n + t) = \begin{cases} Y - \frac{1}{m_2}t, & 0 < t < m_2 Y, \\ 0, & m_2 Y < t < a_1 Y, \end{cases} \tag{5.7}$$

$$\bar{Q}_3(T_n + t) = \begin{cases} \left(\frac{1}{m_2} - \frac{1}{a_1}\right)t, & 0 < t < m_2 Y, \\ Y - \frac{1}{a_1}t, & m_2 Y < t < a_1 Y. \end{cases} \tag{5.8}$$

Fig 5 illustrates the fluid scaled sample paths as $m_1 \searrow 0$. We plot $\bar{Q}_2(t)$ and $\bar{Q}_2(t) + \bar{Q}_3(t)$ against t.

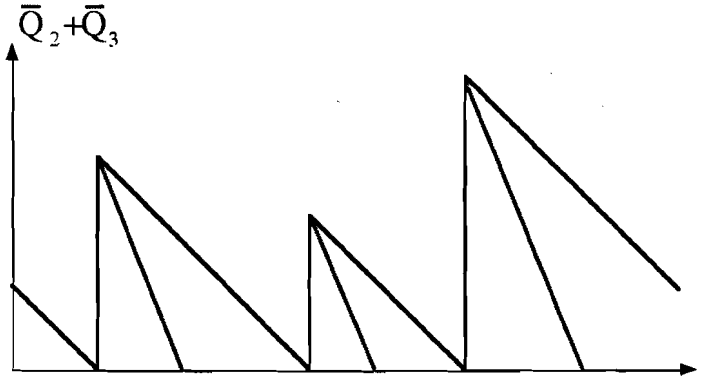

Figure 5: The fluid scaled sample path of $\bar{Q}_2, \bar{Q}_2 + \bar{Q}_3$ for small m_1

Proof. To prove (i) we express α_2, α_3 in terms of a_1, m_2, m_3, and calculate derivatives with respect to m_3. After straightforward though lengthy calculus and algebraic manipulations, in which we use $a_1 > m_2$, we obtain that these derivatives are > 0. Substituting $m_3 = 0$ we get $\alpha_2 = \frac{m_2}{a_1}$, $\alpha_3 = 0$ which proves (ii).

We provide the detailed proof of parts (i), (ii) in the Appendix

We now prove (iii): We consider small m_1 and $m_3 \approx a_1 > m_2$. We wish to analyze a single cycle, starting from $Q_2 = 1, Q_3 = 0$, and until we get back to $Q_2 = Q_3 = 0$. First there is a push period of length $\sim \exp(\mu_2)$ in which a large number of jobs arrive from buffer 1 to buffer 2. At the end of this time buffer 3 has a single part in it. Buffer 3 now behaves like an unstable M/M/1 queue, with input rate μ_2 and service rate μ_3. Such an M/M/1 queue has a geometric number of finite busy periods, followed by an infinite busy period. Let K be the total number of busy periods, including the infinite one. It is known that the probability that a busy period will be finite (the random walk will return to 0) is $\frac{m_2}{m_3}$. Hence, $\mathbb{P}(K = k) = \left(1 - \frac{m_2}{m_3}\right)\left(\frac{m_2}{m_3}\right)^{k-1}$.

The number of the finite busy periods, and their total duration is independent of m_1. If m_1 is small enough then with very high probability buffer 2 will not be empty during all of

these busy periods.

Assuming this is the case, arrivals from buffer 1 to buffer 2 occur during the K buffer 2 processing intervals, each of which has duration $\sim \exp(\mu_2)$. The sum of this geometric number of exponential durations is $\sim \exp(\mu_2 - \mu_3)$. Hence N, the total number of arrivals in the whole cycle has $\mathbb{P}(N = n) = \left(1 - \frac{\mu_1}{\mu_1+\mu_2-\mu_3}\right)\left(\frac{\mu_1}{\mu_1+\mu_2-\mu_3}\right)^{n-1}$, with expected number $E(N) = \frac{\mu_1+\mu_2-\mu_3}{\mu_2-\mu_3}$.

We now scale time by $1/m_1$, and let $m_1 \to 0$. Then the scaled duration in which input to buffer 2 occurs is 0. The scaled total number of arrivals is $Y \sim \exp(\mu_2 - \mu_3)$. In the remaining duration of the cycle buffer 2 drains to 0 along the straight line $Y - \frac{1}{m_2}t$, and into buffer 3. Fluid outflow from buffer 3 is slower, at rate $\frac{1}{m_3} \approx \frac{1}{a_1}$. ■

References

[1] Chen, R-R., Meyn, S.P. (1999) Value iteration and optimization of multiclass queueing networks, *Queueing Systems Theory and Applications*, 32:65–97.

[2] Cohen, J.W. (1982) *The Single Server Queue*, North Holland, Amsterdam.

[3] Dai, J. G. and Weiss, G. (1994). Stability and Instability of fluid models for re-entrant lines. *Mathematics of Operations Research* **21**, 115–134.

[4] Harrison, J.M. (1988) Brownian models of queueing networks with heterogeneous customer populations. *Proceedings of the IMA Workshop on Stochastic Differential Systems*, Fleming W., Lions P.L., editors, Springer-Verlag.

[5] Kleinrock L. (1975) *Queueing Systems, Vol. I: Theory*, Wiley, New York.

[6] Kumar, P.R. (1993) Re-entrant lines. *Queueing Systems: Theory and Applications* **13**, 87-110.

[7] Kumar S. and Kumar P. R. (1994) Performance Bounds for Queueing Networks and Scheduling Policies, *IEEE Transactions on Automatic Control*, **38**, 1600-1611.

[8] Nazarathy, Y. (2001) *Evaluation of on-line scheduling rules for high volume job shop problems, a simulation study.* M.A. Thesis, University of Haifa.

[9] Ross, S.M. (1983) *Stochastic Processes*, Wiley, New York.

[10] Weiss, G. (1995) On optimal draining of a fluid re-entrant line. *Proceedings of the IMA Workshop on Stochastic Networks*, Kelly, F.P. and Williams, R.J, editors, Springer-Verlag, 1995.

[11] Weiss, G. (1999) Scheduling and control of manufacturing systems — a fluid approach *Proceedings of the 37 Allerton Conference, 21–24 September, 1999, Monticello, Illinois*, 577-586.

[12] Weiss, G. (2003) Stability of a simple re-entrant line with infinite supply of work — the case of exponential processing times. *Journal of the Operations Research Society of Japan*, to appear.

Appendix: Proofs of Propositions 5.1 and 5.2

Proof of Proposition 5.1. From (3.29) we have

$$\lim_{t\to\infty} \mathbb{P}(Q_3(t) > n_3) = \frac{m_3}{m_1+m_3}\alpha_3^{n_3}, \quad n_3 = 0,1,2,\ldots.$$

We need to show that α_3 is increasing with m_2, that it converges to 0 when $m_2 \searrow 0$, and it equals $\frac{m_3}{m_1+m_3}$ when $m_2 = m_1 + m_3$. We take the derivative of α_3 w.r.t μ_2:

$$\frac{d\,\alpha_3}{d\,\mu_2} = \frac{1}{2\mu_3}\left(1 - \frac{\mu_1+\mu_2+\mu_3}{\sqrt{(\mu_1+\mu_2+\mu_3)^2 - 4\mu_1\mu_3}}\right) < 0,$$

which proves that α_3 increases in $m_2 = 1/\mu_2$.

We substitute $\mu_2 = 1/m_2$ to get:

$$\alpha_3 = \frac{\mu_1 m_2 + 1 + \mu_3 m_2 - \sqrt{(\mu_1 m_2 + 1 + \mu_3 m_2)^2 - 4\mu_1\mu_3 m_2^2}}{2\mu_3 m_2}.$$

This is seen to converge to zero as $m_2 \to 0$ by L'hospital's rule.

Substituting $m_2 = m_1 + m_3$ in α_3 we get, after straightforward calculations, that $\alpha_3 = \frac{m_3}{m_1+m_3}$.

From (3.28) we have

$$\lim_{t\to\infty} \mathbb{P}(Q_2(t) > n_2) = \frac{m_2}{m_1+m_3}\alpha_2^{n_2}, \quad n_2 = 0,1,2,\ldots.$$

To show that this is increasing in m_2 for every value of n_2, we show that α_2 is increasing with m_2. We will also show that α_2 equals 0 when $m_2 = 0$, and it converges to 1 when $m_2 \nearrow m_1 + m_3$.

We take the derivative of α_2 w.r.t μ_2:

$$\frac{d\,\alpha_2}{d\,\mu_2} = \frac{\mu_1}{2\mu_3}\left[\frac{1}{\mu_2}\left(-1 + \frac{\mu 1+\mu_2+\mu_3}{\sqrt{(\mu_1+\mu_2+\mu_3)^2 - 4\mu_1\mu_3}}\right) - \frac{1}{{\mu_2}^2}\left(-\mu_1-\mu_2+\mu_3+\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}\right)\right].$$

Multiplying this by $\frac{2\mu_2^2\mu_3}{\mu_1}\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}$, which is a positive quantity, we get, after collecting terms:

$$-(\mu_1-\mu_3)^2 - \mu_2(\mu_1+\mu_3) + (\mu_1-\mu_3)\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}\,.$$

This is obviously negative if $\mu_1 < \mu_3$. If $\mu_1 > \mu_3$ we have a difference of a positive and negative term. Taking squares of both terms we get:

$$\left(-(\mu_1-\mu_3)^2 - \mu_2(\mu_1+\mu_3)\right)^2 - \left((\mu_1-\mu_3)\sqrt{(\mu_1+\mu_2+\mu_3)^2-4\mu_1\mu_3}\right) = 4\mu_1{\mu_2}^2\mu_3 > 0,$$

so the previous expression is negative also when $\mu_1 > \mu_3$. Hence α_2 is decreasing in μ_2, and so it is increasing in m_2.

We substitute $\mu_2 = 1/m_2$ in α_2 to get:

$$\alpha_2 = \frac{\mu_1}{2\mu_3}\left(-\mu_1 m_2 - 1 + \mu_3 m_2 + \sqrt{(\mu_1 m_2 + 1 + \mu_3 m_2)^2 - 4\mu_1\mu_3 m_2^2}\right),$$

and we see immediately that, when $m_2 = 0$, we get $\alpha_2 = 0$. Substituting $m_2 = m_1 + m_3 = \frac{1}{\mu_1} + \frac{1}{\mu_3}$ we get after elementary calculations that $\alpha_2 = 1$. ■

Proof of Proposition 5.2, parts (i) and (ii). To prove (i) we will again show that α_2, α_3 are increasing as m_3 increases. We write α_3 in terms of a_1, m_2, m_3:

$$\alpha_3 = \frac{1}{2}\left(\frac{m_3}{m_2} + \frac{a_1}{a_1 - m_3} + \sqrt{\left(\frac{m_3}{m_2} + \frac{a_1}{a_1 - m_3}\right)^2 - \frac{4m_3}{a_1 - m_3}}\right).$$

We take a derivative of this w.r.t. m_3, to obtain:

$$\frac{d\,\alpha_3}{d\,m_3} = \frac{1}{2}\left(\frac{1}{m_2} + \frac{a_1}{(a_1 - m_3)^2} - \frac{-\frac{4a_1}{(a_1-m_3)^2} + 2\left(\frac{1}{m_2} + \frac{a_1}{(a_1-m_3)^2}\right)\left(\frac{m_3}{m_2} + \frac{a_1}{a_1-m_3}\right)}{2\sqrt{\left(\frac{m_3}{m_2} + \frac{a_1}{a_1-m_3}\right)^2 - \frac{4m_3}{a_1-m_3}}}\right).$$

We now show this is positive. We multiply this by the positive square root term, and obtain positive term times a square root minus a second term. If the second term is negative, we are done. If it is positive, then the whole expression is positive only if the difference of the two squares is positive. We write that out now:

$$\begin{aligned}
&\left[\frac{1}{m_2} + \frac{a_1}{(a_1 - m_3)^2}\right]^2 4\left[\left(\frac{m_3}{m_2} + \frac{a_1}{a_1 - m_3}\right)^2 - \frac{4m_3}{a_1 - m_3}\right] \\
&- \left[-\frac{4a_1}{(a_1 - m_3)^2} + 2\left(\frac{1}{m_2} + \frac{a_1}{(a_1 - m_3)^2}\right)\left(\frac{m_3}{m_2} + \frac{a_1}{a_1 - m_3}\right)\right]^2 = \\
&= \frac{16\left(a_1^3 m_2 - 2a_1^2 m_2 m_3 + a_1^2 m_3^2 + 2a_1 m_2 m_3^2 - 2a_1 m_3^3 + m_3^4\right)}{m_2^2 (a_1 - m_3)^4}.
\end{aligned}$$

This has a positive denominator. We consider the numerator. We cancel the 16 and substitute $a_1 = m_1 + m_3$, and obtain:

$$m_1^3 m_2 + m_1^2 m_2 m_3 + m_1^2 m_3^2 + m_1 m_2 m_3^2 + m_2 m_3^3,$$

which is positive.

We turn to α_2. We write α_2 in terms of a_1, m_2, m_3:

$$\alpha_2 = \frac{m_2}{2(a_1 - m_3)}\left(-\frac{m_3}{m_2} + \frac{a_1 - 2m_3}{a_1 - m_3} + \sqrt{\left(\frac{m_3}{m_2} + \frac{a_1}{a_1 - m_3}\right)^2 - \frac{4m_3}{a_1 - m_3}}\right).$$

We take the derivative of α_2 w.r.t. m_3:

$$\frac{d\alpha_2}{dm_3} = \frac{1}{2(a_1-m_3)^3}\left(-a_1^2 + a_1 m_3 - 2m_2 m_3 + \right.$$
$$\left. + \frac{a_1^3 m_2 + a_1^3 m_3 + a_1^2 m_2 m_3 - 2a_1^2 m_3^2 - 2a_1 m_2^2 m_3 - 2a_1 m_2 m_3^2 + a_1 m_3^3 + 4m_2^2 m_3^2}{\sqrt{a_1^2 m_2^2 + 2a_1^2 m_2 m_3 + a_1^2 m_3^2 - 4a_1 m_2^2 m_3 - 2a_1 m_2 m_3^2 - 2a_1 m_3^3 + 4m_2^2 m_3^2 + m_3^4}}\right).$$

Rewriting this when we substitute back $a_1 = m_1 + m_3$ we obtain:

$$\frac{d\alpha_2}{dm_3} = \frac{1}{2m_1^3}\left(-m_1^2 - m_1 m_3 - 2m_2 m_3 + \right.$$
$$\left. + \frac{m_1^3 m_2 + m_1^3 m_3 + 4m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1 m_2^2 m_3 + 3m_1 m_2 m_3^2 + 2m_2^2 m_3^2}{\sqrt{m_1^2 m_2^2 + 2m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1 m_2^2 m_3 + 2m_1 m_2 m_3^2 + m_2^2 m_3^2}}\right).$$

We will now show that this is positive, using $m_1 + m_3 > m_2$. The quantity in the parenthesis consists of a negative term plus a ratio of two terms, with a positive square root in the denominator. We first show that the numerator of the ratio is positive:

$$m_1^3 m_2 + m_1^3 m_3 + 4m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1 m_2^2 m_3 + 3m_1 m_2 m_3^2 + 2m_2^2 m_3^2 \geq$$
$$m_1^3 m_2 + m_1^3 m_3 + 4m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1(m_1+m_3) m_2 m_3 + 3m_1 m_2 m_3^2 + 2m_2^2 m_3^2 =$$
$$m_1^3 m_2 + m_1^3 m_3 + 2m_1^2 m_2 m_3 + m_1^2 m_3^2 + m_1 m_2 m_3^2 + 2m_2^2 m_3^2 0.$$

where in the first inequality we changed the only negative term, using $-2m_1 m_2^2 m_3 \geq -2m_2(m_1+m_3)m_1 m_3$.

It remains to take the square of this numerator minus the square of the first term times the square of the square root, and show it is positive:

$$\left(m_1^3 m_2 + m_1^3 m_3 + 4m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1 m_2^2 m_3 + 3m_1 m_2 m_3^2 + 2m_2^2 m_3^2\right)^2 -$$
$$\left(-m_1^2 - m_1 m_3 - 2m_2 m_3\right)^2 \left(m_1^2 m_2^2 + 2m_1^2 m_2 m_3 + m_1^2 m_3^2 - 2m_1 m_2^2 m_3 + 2m_1 m_2 m_3^2 + m_2^2 m_3^2\right) =$$
$$-2m_2^2(m_1+m_3) + m_1 m_3(m_1+m_3) + m_2(m_1+m_3)^2 + m_2(m_1^2 + m_1 m_3 + m_3^2) \geq$$
$$-2m_2^2(m_1+m_3) + m_1 m_2 m_3 + m_2(m_1+m_3)^2 + m_2(m_1^2 + m_1 m_3 + m_3^2) =$$
$$-2m_2^2(m_1+m_3) + 2m_2(m_1+m_3)^2 > 0.$$

Next we check (ii): We can write α_3 as

$$\alpha_3 = \frac{1}{2}\left(1 + \mu_1 m_3 + \mu_2 m_3 - \sqrt{(1+\mu_1 m_3 + \mu_2 m_3)^2 - 4\mu_1 m_3}\right)$$

and one can then see immediately that for $m_3 = 0$, $\alpha_3 = 0$.

We can write α_2 as

$$\alpha_2 = \frac{1}{2}\left(1 - \mu_1 m_3 - \mu_2 m_3 - \sqrt{(1+\mu_1 m_3 + \mu_2 m_3)^2 - 4\mu_1 m_3}\right)$$

and one can then see immediately that for $m_3 = 0$, $\alpha_2 = \frac{m_2}{m_1} = \frac{m_2}{a_1}$.

■