Push and Pull Systems With External Demands

Michael Zazanis
Dept. of IEOR
University of Massachusetts
Amherst, MA 01003

September 1994

Abstract

We examine Push and Pull production control systems under a make-to-order policy with safety stock. We compute the distribution of the waiting time until a demand is satisfied, as well as flow time distributions and service levels for both systems. Work-in-process levels are determined as well. The analysis is carried out under Markovian assumptions and the explicit results on flow times depend on non-overtake conditions that are satisfied for single machine stations.

1 Introduction

This paper is a first step in an attempt to model the performance of various Push and Pull production strategies and the way they respond to external demands. These demands could represent either outside orders or signals from downstream cells in the same plant. The production line is modelled as a series of single machine stations with exponential processing times, external demands are Poisson, and, in this paper, two production control schemes are examined: A Push scheme with safety stock S and a Pull scheme with limited Work-in-process and safety stock S.

In the Push scheme each time a demand arrives it immediately authorizes the release of a new job. The demand is either immediately satisfied from the stock or it is backloged. The Pull scheme examined is similar to CONWIP (see [12, 11, 13]). The main difference is that instead of the requirement that the WIP in the system remain constant we require that the total work-in-process including the Finished Goods Inventory (FGI) remain constant and equal to S. External demands, if not immediately satisfied, are again backloged. In this scheme the arrival of a demand authorizes the release of a new job only when the work in process is less than S or equivalently the finished goods inventory is greater than zero. Proper operation of

this scheme, which of course depends crucially upon the right choice for S, results in the same benefits in terms of increased system control as other pull systems (e.g. see [2, 6, 10, 9, 8, 12, 13]).

A number of performance criteria are considered, namely average WIP, the probability that a demand will be backloged, and the mean time to satisfy a demand. We also examine mean flow time as well as flow time variability and obtain explicit expressions in terms of the parameters of the system.

2 Systems under a Push Policy with Safety Stock

In this section, we examine in detail the push policy described above and obtain explicit expressions for its performance. The system consists of M single–machine stations in tandem, the processing times are exponential (with rates μ_i, $i = 1, \ldots, M$), and the external demands are Poisson with rate λ. We will assume that the system is stable i.e. $\rho_i \stackrel{\text{def}}{=} \lambda/\mu_i < 1$ for all i. Finished jobs wait in the FGI buffer and, when a demand arrives, a new job will always be released to join the queue at the first station. Furthermore, if the FGI buffer is not empty, the demand is satisfied immediately and a finished job is removed from the system, otherwise the demand is backloged. We assume that initially there are S finished jobs in FGI (the safety stock) and no WIP. With this initial condition the work in process, (i.e. the unfinished jobs in the M stations) together with the finished jobs in the buffer is always greater than or equal to S.

Consider now the stationary version of this process. Restricting our attention to the M stations (and disregarding FGI), this system behaves as an open Jackson network. Let X_t^i be the WIP at machine i at time t and define

$$X_t = \sum_{i=1}^{M} X_t^i, \tag{1}$$

the total WIP in the system. Since the system is stationary,

$$P(X_0^1 = n_1, \cdots, X_0^M = n_M) = \prod_{i=1}^{M} (1 - \rho_i) {\rho_i}^{n_i}. \tag{2}$$

The negative part of $X_0 - S$, $Y_0 = (X_0 - S)^-$, is the finished goods inventory while the positive part, $Z_0 = (X_0 - S)^+$ is the number of backloged demands. Let τ denote the mean time to fill a demand in steady state. Applying Little's law to the finished goods inventory buffer, we obtain

$$\lambda\tau = E(X_0 - S)^+.$$

We proceed to compute τ explicitly in terms of the parameters of the system:

$$P(X_0 = k) \;=\; \sum_{\{\vec{n}\,:\, n_1 + \cdots + n_M = k\}} \prod_{i=1}^{M} \rho_i^{n_i}(1 - \rho_i) \;=\; G(M, k) \prod_{i=1}^{M} (1 - \rho_i),$$

with $G(M,k)$ given by:

$$G(M,k) \;=\; \sum_{\{\vec{n}\,:\,n_1+\cdots+n_M=k\}} \prod_{i=1}^{M} \rho_i^{n_i}.$$

Therefore,

$$\begin{aligned} E(X_t - S)^- &= \sum_{k=0}^{S} (S-k) P(X_0 = k) \\ &= \sum_{k=0}^{S} (S-k) G(M,k) \prod_{i=1}^{M} (1-\rho_i) \\ &= \prod_{i=1}^{M} (1-\rho_i) \sum_{k=0}^{S} (S-k) G(M,k)\ . \end{aligned}$$

Since $(X_0 - S)^+ = X_0 - S + (X_0 - S)^-$ and $EX_0 = \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}$, we get

$$\tau \;=\; \frac{1}{\lambda} \left\{ \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i} \;-\; S \;+\; \prod_{i=1}^{M} (1-\rho_i) \sum_{k=0}^{S} (S-k) G(M,k) \right\}.$$

Anticipating (7) which is established in the next section,

$$\tau \;=\; \frac{1}{\lambda} \left\{ \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i} \;-\; S \;+\; \prod_{i=1}^{M} (1-\rho_i) \sum_{m=1}^{M} \frac{S - (S+1)\rho_m + \rho_m^{S+1}}{(1-\rho_m)^2 \prod_{l \neq m} (1 - \rho_l/\rho_m)} \right\}. \quad (3)$$

2.1 Customer service criteria

In this framework we can address a number of related issues pertaining to the level of service: The probability that a demand will be satisfied immediately is given by $P(S - X_0 > 0) = P(X_0 < S)$. In view of (1) and (2), the distribution of X_0 is the convolution of N independent geometric random variables (because of the independence of the number of customers in each station). To simplify the analysis we examine only the case where all stations have different utilizations. Then

$$P(X_0 = k) \;=\; \sum_{i=1}^{M} \rho_i^k \, \frac{\prod_{l=1}^{M} (1-\rho_l)}{\prod_{l \neq i} (1 - \rho_l/\rho_i)}. \quad (4)$$

The derivation of (4) is interesting since it does not make use of convolutions directly. The partial fractions expansion through which it is obtained is shown here in the case where $\rho_i \neq \rho_j$ for $i \neq j$. The z-transform of X_0 can then be written as

$$Ez^{X_0} \;=\; \prod_{i=1}^{M} \frac{1-\rho_i}{1 - z\rho_i} \;=\; \sum_{i=1}^{M} \frac{A_i}{1 - z\rho_i}$$

or

$$\prod_{i=1}^{M} (1-\rho_i) \;=\; \sum_{i=1}^{M} A_i \prod_{l \neq i} (1 - z\rho_l)\ .$$

Letting $z = \rho_i^{-1}$ gives

$$A_i = \frac{\prod_{l=1}^{M}(1-\rho_l)}{\prod_{l\neq i}(1-\rho_l/\rho_i)}, \qquad i = 1, 2, \ldots, M.$$

Hence,

$$E[z^{X_0}] \;=\; \sum_{i=1}^{M} \frac{1}{1-z\rho_i} \frac{\prod_{l=1}^{M}(1-\rho_l)}{\prod_{l\neq i}(1-\rho_l/\rho_i)}, \tag{5}$$

from which we obtain (4).

An alternative for computing the distribution of X_0 uses the normalization constants for CQNs for which a number of efficient computational algorithms exist.

$$\begin{aligned} P(X_0 = k) &= \sum_{\vec{n}:n_1+\cdots+n_M=k} \prod_{i=1}^{M}(1-\rho_i)\rho_i^{n_i} \\ &= G(M,k)\prod_{i=1}^{M}(1-\rho_i)\;. \end{aligned} \tag{6}$$

A comparison between (4) and (6) suggests the following expression for the normalization constant in a CQN

$$G(M,k) \;=\; \sum_{i=1}^{M} \frac{\rho_i^k}{\prod_{l\neq i}(1-\rho_l/\rho_i)}. \tag{7}$$

The above closed form expression for the normalization constant was first obtained in [7] (see also [5]).

2.2 Probability that a demand will be satisfied immediately

The probability that a demand will be satisfied immediately is given by

$$\begin{aligned} P(X_0 < S) &= \sum_{k=0}^{S-1} P(X_0 = k) \\ &= \prod_{i=1}^{M}(1-\rho_i) \sum_{k=0}^{S-1} G(M,k) \\ &= \sum_{i=1}^{M}\sum_{k=0}^{S-1} \rho_i^k \frac{\prod_{l=1}^{M}(1-\rho_l)}{\prod_{l\neq i}(1-\rho_l/\rho_i)} \\ &= \sum_{i=1}^{M}(1-\rho_i^S) \prod_{l\neq i} \frac{1-\rho_l}{1-\rho_l/\rho_i}. \end{aligned} \tag{8}$$

2.3 Total WIP in an open system

From the above analysis we can easily obtain an expression for the distribution of the total WIP in an open system:

$$P(X_0 \leq n) \;=\; \prod_{i=1}^{M}(1-\rho_i)\sum_{k=0}^{n} G(M,k) \;=\; G(M+1,n)\prod_{i=1}^{M}(1-\rho_i),$$

where $G(M+1,n)$ is the normalization constant of a CQN with n customers and $M+1$ stations with mean service times $\rho_1,\ldots,\rho_M,1$, or equivalently $\mu_1^{-1},\ldots,\mu_M^{-1},\lambda^{-1}$. In view of (7)

$$G(M,n) \;=\; \sum_{m=1}^{M} \frac{\rho_m^n}{(1-1/\rho_m)\prod_{l\neq m}(1-\rho_l/\rho_m)} \;+\; \frac{1}{\prod_{m=1}^{M}(1-\rho_l)} \tag{9}$$

and hence

$$P(X_0 \leq n) \;=\; 1 \;-\; \sum_{m=1}^{M} \frac{\rho_m^n}{(1/\rho_m-1)\displaystyle\prod_{l\neq m}(1-\rho_l/\rho_m)}\,.$$

3 Limited WIP systems with safety stock

We now consider a make–to–order system with safety stock under a Limited WIP policy. As before the cell consists of M single–machine stations (exponential processing times with rates μ_i). External (or down stream) demands are Poisson with rate λ and, when the system starts, there are S finished parts in Finished Goods Inventory (FGI) and no parts are being processed. Denote by X_t^i the number of parts in process or waiting in front of machine i at time t, and by $X_t = \sum_{i=1}^{M} X_t^i$ the total WIP. When a demand arrives, if there is a part in FGI then it is satisfied immediately and a new part is authorized to start processing at (or join the queue in front of) station 1. If however upon the arrival of the demand no parts are available in FGI then the demand is backloged until a part is finished. Backloged demands are satisfied on a FCFS basis. Furthermore, *a backloged demand does not authorize the production of a new part.* Let Y_t denote the number of parts in FGI with negative values of Y_t corresponding to the number of backloged orders. Under the above policy, $X_t \leq S$ with equality holding whenever $Y_t \leq 0$ (i.e. whenever there are no parts in FGI but, possibly, a demand backlog).

To analyze the performance of this policy we consider now a system, equivalent to the one just described under our Markovian assumptions. It consists of an *open* tandem network with M single–server exponential stations (with the same processing rates). This system has a global buffer of size S (i.e. the total number of customers in the network, X_t, is constrained to be $\leq S$. Customers arrive from outside according to a Poisson process with rate λ and if upon arrival the number of parts in the network is equal to S then they wait in front of the global buffer according to a FCFS discipline. As soon as a customer finishes processing at station M, a customer waiting outside the global buffer is immediately admitted and joins the queue, or starts processing, at station 1. Let Z_t be the total number of customers at time t, including those waiting outside the buffer. When $Z_t \leq S$ in this equivalent system, the original system has no backloged demands, Z_t parts in process (WIP) and $S - Z_t$ parts in FGI. When $Z_t > S$, there are S parts in process (the maximum allowed), $Z_t - S$ unsatisfied demands, and of course no parts in FGI.

In the next section we provide some results concerning the distribution of inter-output times in cyclic, single–server CQNs, which will be necessary in analyzing the

performance of the above policy.

3.1 The interoutput distribution for cyclic CQNs

Consider a Closed Queueing Network consisting of M single server stations with processing rates μ_i and k customers. Let T_n denote the point process of successive outputs from the last station (i.e. the process of finished parts) and $(X_t^1, \ldots, X_t^M)$ (defined as a process with *right–continuous paths* the state of the system. Suppose that the system is stationary under P and denote by P^* the Palm transformation of P with respect to $\{T_n\}$. E^* denotes expectation w.r.t. P^*. We will compute $E^*[e^{-sT_1}]$ and $E[e^{-sT_1}]$. (Of course, $E[e^{-sT_1}] = \frac{1-E^*[e^{-sT_1}]}{sE^*[T_1]}$.) The first Laplace transform can be obtained from a straight-forward application of the arrival theorem. Indeed, if at time $t = 0$ a customer has just left station 1, then

$$E^*[e^{-sT_1}|X_0^M > 0] = \frac{\mu_M}{\mu_M + s},$$

and, generally,

$$E^*[e^{-sT_1}|X_0^M = 0, X_0^{M-1} = 0, \ldots, X_0^i > 0] = \prod_{j=i}^{M} \frac{\mu_j}{\mu_j + s}.$$

The above argument simply takes into account the possibility that the last $M - i + 1$ stations may be idle. From the arrival theorem,

$$P^*(X_0^1 > 0) = \frac{\rho_1 G(M, k-2)}{G(M, k-1)},$$

and

$$P^*(X_0^M = 0, \ldots, X_0^{i+1} = 0, X_i > 0) = \frac{\rho_i G(i, k-2)}{G(M, k-1)}.$$

Hence,

$$E^0[e^{-sT_1}] \;=\; \sum_{i=1}^{M} \frac{\rho_i G(i, k-2)}{G(M, k-1)} \prod_{j=i}^{M} \frac{\mu_j}{\mu_j + s}.$$

The same argument can be used to obtain the expression

$$E[e^{-sT_1}] \;=\; \sum_{i=1}^{M} \frac{\rho_i G(i, k-1)}{G(M, k)} \prod_{j=i}^{M} \frac{\mu_j}{\mu_j + s}$$

directly. In particular,

$$\alpha_k \stackrel{\text{def}}{=} E^*T_1 = \sum_{i=1}^{M} \frac{\rho_i G(i, k-2)}{G(M, k-1)} \left(\frac{1}{\mu_i} + \cdots + \frac{1}{\mu_M} \right) , \tag{10}$$

and

$$\beta_k \stackrel{\text{def}}{=} ET_1 = \sum_{i=1}^{M} \frac{\rho_i G(i, k-1)}{G(M, k)} \left(\frac{1}{\mu_i} + \cdots + \frac{1}{\mu_M} \right) . \tag{11}$$

4 Flow time distributions for Push and Pull systems

The Pull strategy with WIP limited above by N can be modelled in the markovian case as an open queueing network with a global buffer of size N. Demands arrive according to a Poisson process and are admitted to the system only if the total number of customers present, X_t, is less than N. Otherwise they wait outside the global buffer. If at time t $X_t \leq N$ then no demands are backlogged and $N - X_t$ represents finished goods inventory. If, on the other hand, $X_t > N$ then $X_t - N$ represents the number of backlogged demands. If the probability that a demand is backlogged is small the operation of this system can be adequately approximated by a closed queueing network.

For a closed queueing network, from Boxma, Kelly, and Könheim (1984), and Daduna (1982), it follows that, if T_i is the flow time through the i'th station, the joint Laplace transform for the flow times of a tagged customer through the stations satisfies the following product form relationship

$$E[e^{-s_1T_1-\cdots-s_MT_M}] \;=\; \sum_{\vec{j}\in\mathcal{S}(M,N-1)} p(j_1,\ldots,j_M)\prod_{i=1}^{M}\left(\frac{\mu_i}{\mu_i+s_i}\right)^{j_i+1}\;, \tag{12}$$

where $\mathcal{S}(M,N) = \{\vec{j} : j_1+\cdots+j_M = N\}$. (The above holds provided that a non-overtake condition holds, which of course is the case for cyclic single server networks.) Setting $\rho_i = 1/\mu_i$, the rhs of the above equation can be written as

$$\sum_{\vec{j}\in\mathcal{S}(M,N-1)} \frac{\rho_1^{j_1}\cdots\rho_M^{j_M}}{G(M,N-1)}\prod_{i=1}^{M}\left(\frac{\mu_i}{\mu_i+s_i}\right)^{j_i+1}\;,$$

or, equivalently,

$$\prod_{i=1}^{M}\left(\frac{\mu_i}{\mu_i+s_i}\right)\sum_{\vec{j}\in\mathcal{S}(M,N-1)}\prod_{i=1}^{M}\left(\frac{1}{\mu_i+s_i}\right)^{j_i}\frac{1}{G(M,N-1)}\;. \tag{13}$$

The first factor in the above expression corresponds to the joint Laplace transform of the processing times for a job while the second to the joint Laplace transform of waiting times at the stations. Setting $s_1 = \cdots = s_M = s$ in (13) and taking into account (7) we obtain the following expression for the Laplace transform of the cycle time:

$$Ee^{-sT} \;=\; \frac{\displaystyle\sum_{m=1}^{M}\frac{\alpha_m}{(\mu_m+s)^N}}{\displaystyle\sum_{m=1}^{M}\frac{\alpha_m}{\mu_m^N}}\;, \tag{14}$$

where

$$\alpha_m \;=\; \prod_{l\neq m} \frac{1}{\mu_l - \mu_m}\;. \tag{15}$$

From (14) we readily obtain the following expressions for the moments of the cycle time:

$$ET^k \;=\; N(N+1)\cdots(N+k-1)\;\; \frac{\sum_{m=1}^{M} \alpha_m\, \mu_m^{-(N+k)}}{\sum_{m=1}^{M} \alpha_m\, \mu_m^{-N}}\;, \qquad k=1,2,\ldots. \tag{16}$$

The throughput of the closed system with N customers can easily be computed from the first moment of the cycle time (16) and Little's Law:

$$\lambda(N) \;=\; \frac{\sum_{m=1}^{M} \alpha_m\, \mu_m^{-N}}{\sum_{m=1}^{M} \alpha_m\, \mu_m^{-(N+1)}}\;. \tag{17}$$

From (16), (17) we also obtain the following version of Little's Law for the moments of the cycle time in a closed network:

$$\lambda(N)\lambda(N+1)\cdots\lambda(N+k-1)\; ET^k \;=\; N(N+1)\cdots(N+k-1)\;. \tag{18}$$

The coefficient of variation of the cycle time with N customers in the system can be written as

$$C_v(N) \;=\; \sqrt{\left(1+\frac{1}{N}\right)\frac{\lambda(N)}{\lambda(N+1)} - 1}\;. \tag{19}$$

Given that $\lambda(N)$ is increasing in N we have the (asymptotically tight) bound:

$$C_v(N) \leq \sqrt{\frac{1}{N}}\;.$$

For the corresponding open system the Laplace transform of the flow time is of course

$$E[e^{-sT_O}] \;=\; \prod_{m=1}^{M} \frac{\mu_m - \lambda}{\mu_m - \lambda + s}\;, \tag{20}$$

where λ is given by (17). We compare the two systems assuming that they have the same throughput i.e. with $\lambda = \lambda(N)$.

References

[1] Boxma, O.J., F.P. Kelly, and A.G. Konheim (1984). "The Product Form for Sojourn Time Distributions in Cyclic Exponential Queues," *J. ACM,* **31,** 1, 128-133.

[2] Buzacott, J.A. and J.G. Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*, Prentice Hall, 490-515.

[3] Daduna, H., (1982). "Passage times for overtake-free paths in Gordon-Newell networks," *Adv. Appl. Probab.* **14,** 672-686.

[4] Delersnyder, J.-L., T.J. Dodgson, H. Muller, and P.J. O'Grady (1989). "Kanban controlled pull systems: an analytic approach", *Management Science,* **35,** 1079-1091.

[5] Gordon, J.J. (1990). "The Evaluation of Normalizing Constants in Closed Queueing Networks," *Opns. Res.* **38,** 5, 863-869.

[6] Hall, R. W. (1983). *Zero Inventories*, Dow Jones-Irwin, Homewood, Ill.

[7] Harrison, P.G. (1985). "On Normalizing Constants in Queueing Networks," *Opns. Res.* **33,** 464-468.

[8] Karmarkar, U. S. (1986). "Push, Pull, and Hybrid Control Schemes", Working Paper Series No. QM8614, Graduate School of Management, University of Rochester, Rochester, New York.

[9] Mitra, D., and I. Mitrani (1990). "Analysis of a Kanban discipline for cell coordination in production lines, I", *Management Science,* **36,** 1548-1566.

[10] Mitra, D., and I. Mitrani (1991). "Analysis of a Kanban Discipline for Cell Coordination in Production Lines, II: Stochastic Demands", *Operations research*, Vol. 39, No. 5, 807-823.

[11] Spearman, M. (1992). "Customer Services in Pull Production Systems", *Operations research*, Vol. 40, No. 5, 948-958.

[12] Spearman, M., D. Woodruff, and W. Hopp (1990). "CONWIP: a Pull Alternative to Kanban ", *Int. j. Prod. Res.*, **28**, 5, 879-894.

[13] Spearman, M., and M. Zazanis (1992). "Push and Pull Production Systems: Issues and Comparisons", *Operations research*, Vol. 40, No. 3, 521-531.