The impact of arrival rate control on the economic performance of make-to-order firms with work-in-process (in)dependent order processing rates

THE IMPACT OF ARRIVAL RATE CONTROL ON THE ECONOMIC PERFORMANCE OF MAKE-TO-ORDER FIMRS WITH WORK-IN-PROCESS (IN) DEPENDENT ORDER PROCESSING RATES.

H.P.G. van Ooijen and J.W.M. Bertrand
Department of Technology Management
Technische Universiteit Eindhoven

December 2003

THE IMPACT OF ARRIVAL RATE CONTROL ON THE ECONOMIC PERFORMANCE FOR MAKE-TO-ORDER FIRMS WITH WORK-IN-PROCESS (IN) DEPENDENT ORDER PROCESSING RATES.

Abstract

Many production systems operate, for economical reasons, at a rather high average utilization rate. Given the dynamics in order arrival this often implies operating with a high work-in-process. If there are costs associated with the level of work-in-process this may lead to high costs. So, in general, a trade-off is needed between the income (determined by the production throughput) and the costs of work-in-process. In this paper we investigate the impact on operational performance of the production system of being able to influence the order arrival rate. Operational performance is measured as income due to products sold minus costs incurred by work-in-process and changing the order arrival rate. The results indicate that for work-in-process independent order processing rates, depending on the costs of switching the arrival rate and depending on the work-in-process carrying costs, arrival rate control can result in 1.5 to 3% improvement in operational performance. For work-in-process dependent order processing rates, as modeled in this research, the analysis of our models suggest that improvements in operational performance are possible in a range between 9 and 16%.

Keywords: arrival rate switching, operational performance, work-in-process dependent processing rates

1. Introduction

Many production systems operate, for economical reasons, at a rather high average utilization rate. Given the dynamics in order arrival this often implies operating with a high work-in-process. If there are costs associated with the level of work-in-process this may lead to high costs. So, in general, a trade-off is needed between the income (determined by the production throughput) and the costs of work-in-process.

If the work-in-process carrying costs are high it may be beneficial to avoid high levels of work-in-process. This can be done by limiting the level of work-in-process (the CONWIP approach (Hopp and Spearman [2000])), by operating at low order arrival rate, or by decreasing the order arrival rate if the level of work-in-process is too high and increasing to the order arrival rate if the level of work-in-process is too low. Each of these options has its disadvantages. Limiting the level of work-in-process means that no new orders are accepted as long as the level of work-in-process is at its maximum level. This may be unacceptable, since the policy is arbitrary regarding the customer orders that will not be accepted, and customers may be lost that are crucial to the long-term success of the firm. Deliberately operating at a low order arrival rate leads to a situation where the capacity utilization will be low, and production costs will be high.
Increasing and decreasing order arrival rates means that the company has to build up the capability to do so, which implies a certain cost each time that the order arrival rate is increased or decreased. Whether or not a company should use this type of arrival rate control depends on the improvement in performance that can be obtained, including the effects on throughput and work-in-process.

Order arrival rate control might even be more relevant if the order processing rates of the production system depends on the work-in-process. Recently research has been published which provides evidence for the existence of a relationship between work-in-process and productivity (Schmenner [1988], Holström [1994] and Liebermann and Demeester [1999]). Work-in-process is measured as the number of work orders on the shop floor and productivity is measured as output per employee. There seems to be a certain level of work-in-process at which the output is at its maximum. If the work-in-process differs from this level the output per employee decreases and thus the total output decreases. There exists evidence from psychological research that supports this relationship between workload and productivity. The argument builds on the assumption that performing operations requires human perception, human information processing, human decision-making and human actions to take place. When the number of work orders on the shop floor increases, the perceived work pressure increases, which increases the level of arousal of the shop floor personnel. Increase of arousal first leads to an increase in operator efficiency, up to a certain maximum, and then leads to a decreasing operator efficiency (see e.g. Wickens [1992], chapter 10, Stress and Human Error).

Under such circumstances management might want to control the work-in-process in the shop such that it is close to the level where operator efficiency is at its maximum. This can be achieved by deliberately manipulating the order arrival rate, although at a cost. In this paper we investigate the impact on operational performance of a production system of being able to influence the order arrival rate. Operational performance is measured as income due to products sold minus costs incurred by work-in-process and changing the order arrival rate.

The remainder of this paper is as follows. In Section 2 we will discuss the relevant literature. In section 3 we will present the economical setting of the problem, the arrival rate switching policy, and a model of the relationship between work-in-process and order-processing rate. Section 4 discusses the method that has been used to find the optimal switching values. In Section 5 the results for different economical settings are presented and Section 6 discusses these results. The paper is completed with the conclusions in Section 7.

2. Literature review.

In this section we will discuss the relevant literature on determining the optimal arrival rate(s). We will start with the situation where the operator efficiency is independent of the work-in-process.

One of the first to study the arrival rate in an economical setting was Hillier [1963]. He studies the problem of determining the proper balance between the amount of service and the amount of waiting for that service. He uses an economic model for determining the level of service, which minimizes the total of the expected cost of service and the expected cost of waiting for that service.

Tijms and van der Duyn Schouten [1977] study a system that at each point in time can be in one of two possible stages 1 and 2, and where at any moment the system can be switched from one stage to another. The arrival rate and the service rate both depend on the stage of the system. They derive a formula for the long run average expected cost per unit time if holding costs are stage dependent and switch over costs are fixed.

Buss et al. [1994] determine for a single machine job shop-like production system, amongst others, the arrival rate and the capacity level that maximizes the profit if cost related to capacity and to work-in-process are declining to scale. They assume fixed order processing rates.
So and Song [1998] study a production system where demand is sensitive to both price and lead time and determine the joint optimal selection of price, lead time and capacity.

A number of studies have been published which investigate situations where the processing rate depends on the level of work-in-process. These studies can be considered to be more or less symmetric to the situations with work-in-process dependent arrival rates.
Cohen [1976] investigates a single server with two service rates, where the service rate is changed if the level of work-in-process exceeds a certain threshold K, and there are costs related to switching and the level of work-in-process. Purpose of the study is to determine the optimal value for K.
Tijms [1977] studies a server with two constant service rates and fixed switch over costs. The server switches from low service rate to high service rate when the work-in-process exceeds a level y_1 and switches from a high service rate to a low service rate when the work-in-process falls to a level y_2, where $0 \leq y_2 \leq y_1$. Tijms considers a linear holding cost, rate dependent service-cost and fixed switchover costs. He derives an explicit expression for the average cost as a function of y_1 and y_2.
Doshi et al. [1978] study a production-inventory system with two possible production rates, where the control is based on two critical inventory levels, and where the arrival rate and the distribution of the demand depend on the current production rate. This paper gives the long run average costs per unit time as a function of the critical inventory levels.
Nishimura and Jiang [1995] consider a server with two service modes: regular speed and high speed, and a service rule that is characterized by two switchover levels. A key feature is that the server takes vacation for setup before a new service mode is available. The paper derives for the general model an expression for the generating function of the equilibrium queue-length distribution in terms of the switchover levels.
Bar-Lev et al. [1996] analyze a stochastic production/inventory problem with state (i.e. work-in-process) dependent production rates. When the inventory W(t) falls below a critical level m, production is started at a rate of r[W(t)], i.e. the production rate dynamically changes as a

function of the inventory level. Production continues until a level M is reached. The two-sided (m,M) policy is optimized using the expected cost obtained from the stationary distribution of W.

Related to the research on arrival rate switching is the research on admission control. In systems with admission control customers arrive according to a process with a fixed rate; however, a customer, upon arrival, may either be permitted or prohibited to enter the system. Examples of research on admission control are the following. Stidham [1985] reviews open-loop systems where each arriving job is accepted with probability *p*, and closed-loop systems where arriving jobs are admitted or not based on the observed queue length. The main emphases are on the difference between socially optimal and individually optimal (equilibrium) controls and on the use of dynamic programming inductive analysis to show that an optimal control is monotonic or characterized by one or more "critical numbers".
Xu and Shantikumar [1993] determine the optimal admission control policy for a first come, first served *M/M/m* ordered-entry queueing system to maximize the expected discounted profit. They derive an easily computable approximation for the optimal threshold value is derived

Our purpose is to investigate whether it is economically profitable to use an arrival rate switching policy, if changing the arrival rate goes at a cost. As in previous research we consider switching policies that are characterized by two switching levels and two arrival rates. However, all four parameter values have to be determined simultaneously in order to find the combination that maximizes the operational performance of the system. In the next section we describe the economical setting of the problem and present the models of the production system for work-in-process independent order processing rates and work-in-process dependent order processing rates.

3. The economic setting.

We consider a production-to-order firm that operates in a competitive market as a price taker. We assume that the firm operates a symmetrical job shop-like production system where the order processing rate μ_t of the system can be modeled as a function of the number of machines, m, and the number of orders in process, W_t, as follows:

$$\mu_t = \left(\frac{W_t}{m + W_t} \right) \left(\frac{m}{p.g} \right)$$

with g = average number of operations per production order

p = operation processing time

Assuming an average operation processing time of 1 hours, 5 machines, and on average 5 operations per work order, the shop can at maximum produce 1 order per hour. Assuming 50 working weeks of 40 hours a year, a maximum of 2000 orders can be processed per year.

We assume that the arrival rate, λ_t, results from production orders placed by a set of customers under a supply contract per customer. The supply contract states the range of the total amount of products to be delivered over a contract period and leaves freedom for the customer regarding the exact amount and timing of orders to be placed. We assume that the aggregate order arrival process resulting from the orders placed by all the customers under their contracts can be approximated as a Poisson process (assuming many customers).

The firm can influence the aggregate order arrival rate by increasing or decreasing the customer base, or by trying to increase or decrease the supply volume in the contract per customer upon renewal of the contract. We assume that the firm controls the order arrival rate using a hysteretic rule (as in many studies on switching, c.f. Nobel [1998] that works as follows:

- As long as the work-in-process is within a specific range of values it tries to renew contracts such that the existing aggregate order arrival rate is maintained.
- If the work-in-process exceeds a specific value, W_{hl}, it decreases the aggregate order arrival rate to a value λ_l by trying not to renew one or more contracts or decreases the supply volume in one or more contracts.
- If the work-in-process drops below a specific value, W_{lh}, the firm tries to increase the order arrival rate to a value λ_h by attracting new customers or to increase the supply volume in one or more contracts.

Increasing the aggregate order arrival rate goes at a cost, C_{lh}, and decreasing the aggregate order arrival rate goes at a cost, C_{hl}. We further assume that firm receives an average contribution of m per order, and that the firm incurs a cost of work-in-process that is proportional to the average value of the work-in-process. We assume that the material costs can be neglected compared to the value added and that the costs of work-in-process could be modeled as $r.m.W_t$. Let f denote the frequency of increasing or decreasing the arrival rate per unit of time, T the average throughput per unit of time and W the average level of work-in-process. Then the firm wants to determine switching values W_{hl} and W_{lh}, and arrival rates λ_l and λ_h such that the operational performance:

$$P = T.m - f(C_{lh} + C_{hl}) - r.m.W \tag{1}$$

is maximized.

In this paper we investigate what can be the benefits for a manufacturing firm of being able to deliberately increase or decrease the customer order arrival rate. We investigate this for different values of work-in-process carrying costs and different costs of changing the customer order arrival rate. We model the production system as a symmetrical 5 work center job shop with one machine per work center and study two variants of this system. In the first variant that serves as a reference the order-processing rate a given constant, μ. In the second variant, the order processing rate, μ_t, is a function of the number of orders in the shop (or the work-in-process, W_t) with the property that for low work-in-process the processing rate is low, for higher work-in-process values the processing rate increases to a maximum value, μ, and then, for still higher values of work-in-process, decreases again to a low value. This models the effects of perceived work pressure on operator efficiency as found in research in cognitive psychology (Wickens [1992]). Figure 1 shows the curve we used in our research.

4. Determining the optimal parameters

Finding optimal combinations of switching levels and arrival rates, for a given production system, has been achieved with a combination of numerical computations and optimal search technology.

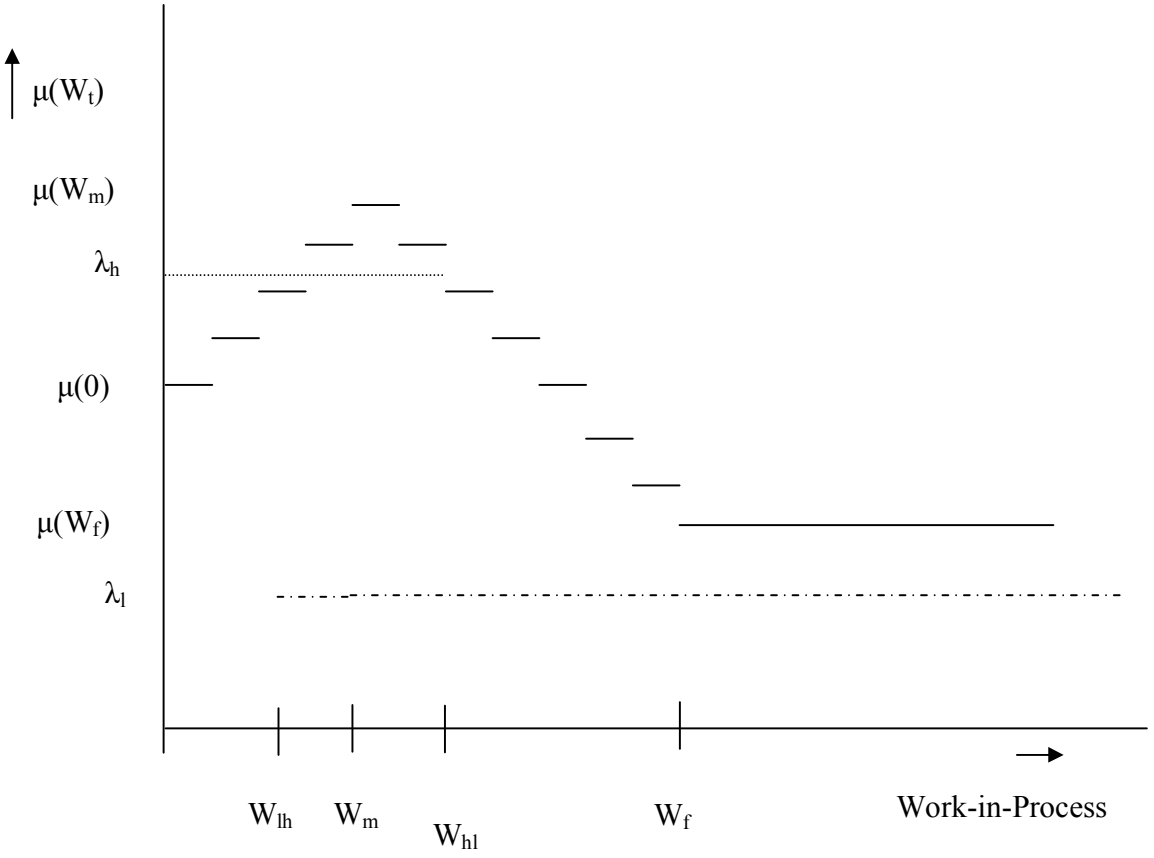

Fig. 1 The order arrival rate and the order processing rate as a function of the number of the orders in the shop

We assume that the departure rate increases from $\mu(0)$ to a maximum value $\mu(W_m)$ in equal steps from $W_t=0$ to $W_t=W_m$, and decreases from $\mu(W_m)$ to $\mu(W_f)$ in equal steps from $W_t = W_{m+1}$ to $W_t = W_f$. For $W_t > W_f$, $\mu = \mu(W_f)$. We further assume that management in the shop is aware of the existence of a dependency of the production efficiency on the workload in the shop and has knowledge of both the maximum departure rate of orders $\mu(W_m)$ and the minimum departure rate, $\mu(W_f)$, and of the work-in-process for which the maximum departure rate is obtained. Figure 1 shows possible values of the decision variables values W_{hl}, W_{lh}, λ_l and λ_h

Stability requires that $\lambda_\ell < \mu(W_f)$. Note that without arrival rate control (thus $\lambda=\lambda_h=\lambda_\ell$) this implies that $\lambda< \mu(W_f)$, which may result in a rather low utilization of capacity.

Our order processing rate function models the inverse U-shape type of dependency of production efficiency on perceived work pressure as reported in Wickens (1992). In our model we thus assume that the work pressure that an operator experiences is proportional to the work-in-process in the shop and that the perceived work pressure affects this production efficiency. It will be clear that the management of the shop would prefer to operate the shop at the highest possible level of efficiency, that is, realize the order processing rate $\mu(W_m)$. However, in our model, management cannot directly influence the order-processing rate since the order-processing rate depends on W_t, which is a stochastic variable. In our model, the parameters that are under directly control of management are the order arrival rates λ_h and λ_ℓ, and the switching values W_{hl} and W_{lh}. Now the question is, given the relationship between order processing rate and work-in-process as in Figure 1, revenue per order equal to *m*, costs related to switching the arrival rate equal to C_{lh} and C_{hl}, and costs of work-in-process equal to $r.m.W_t$, what are then the optimal switching points W_{hl}, W_{lh}, and the optimal arrival rates λ_h, λ_ℓ, and what is then the resulting operational performance per unit time?

To find optimal switching parameter values we modeled the shop as a two dimensional Markov chain by including the arrival rate as a variable in a birth-death queueing process. To calculate the steady state probability of having n orders in the system, P_n, we introduce $P(n,1)$ as being the probability of having n customers in the system when the arrival rate is high, and $P(n,2)$ as being the probability of having n customers in the system when the arrival rate is low. Given the arrival rate and order processing rate as discussed in Section 2, the P_n's can be calculated for any feasible combination of parameter values from the following state equilibrium relationships:

$$P(n,1)[\lambda_h + \mu(n)] = \lambda_h P(n-1,1) + \mu(n+1)P(n+1,1) \qquad n < W_{hl}-1$$

$$P(W_{hl}-1,1)[\lambda_h + \mu(W_{hl}-1)] = \lambda_h P(W_{hl}-2,1)$$

$$P(W_{lh},1)[\lambda_h + \mu(W_{hl})] = \lambda_h P(W_{lh}-1,1) + \mu(W_{lh}+1)P(W_{lh}+1,1) + \mu(W_{lh}+1)P(W_{lh}+1,2)$$

$$P(W_{lh}+1,2)[\lambda_l + \mu(W_{lh}+1)] = \mu(W_{lh}+2)P(W_{lh}+2,2)$$

$$P(n,2)[\lambda_l + \mu(n)] = \lambda_l P(n-1,2) + \mu(n+1)P(n+1,2) \qquad W_{lh}+1 < n$$

$$P(W_{hl},2)[\lambda_l + \mu(W_{hl})] = \lambda_l P(W_{hl}-1,2) + \mu(W_{hl}+1,2) + \lambda_h P(W_{hl}-1,1)$$

$$\sum_{n=1}^{W_{hl}-1} P(n,1) + \sum_{n=W_{lh}+1}^{\infty} P(n,2) = 1$$

From the P_n values the average level of work-in-process (W), the average throughput per unit time (T) and the average switching frequencies per unit time (f) can be calculated numerically. Next the operational performance per unit time can be calculated using (1).

Knowing the operational performance per unit time for given values of W_{hl}, W_{lh}, λ_h and λ_ℓ, the question remains how to find optimal parameter values. This means that we have to maximize (1) over W_{hl}, W_{lh}, λ_h and λ_ℓ. Since T, f and W depend on the P_n*'s* and we do not have explicit expressions for the P_n values, we have used the computer software system OptQuest®. This system automatically searches for the optimal solution for complex models, using a combination of tabu search, scatter search, integer programming and neural networks.

5. The operational performance as a function of the economic parameters

To investigate the effects of arrival rate switching on the optimal operational performance, we calculated the optimal performance for different scenarios with regard to the economic parameters m, r, c_{hl} and c_{lh}. The values that have been used for these parameters are given in Table 1. We scaled the values of the parameters to the revenue per order, m, which therefore has been set equal to 1. The first three scenarios represent situations where switching the arrival rate is excessively costly. The last three scenarios represent situations where switching costs are zero. Both sets of scenarios have been added as references and in order to determine the maximum benefit that can be obtained from being able to switch arrival rates.

We calculated the optimal operational performance with these scenarios for two production situations. In the first production situation the processing rate is independent of the work-in-process and in the second production situation the processing rate depends on the work-in-process according to the model described in Section 3 with $\mu(0)=0.875$, $W_m= 45$, $\mu(W_m) = 1$, $W_f = 135$.

Scenario	M	c_{hl}	c_{lh}	r
1	1	1000000	1000000	0.00001
2	1	1000000	1000000	0.0001
3	1	1000000	1000000	0.001
4	1	100	100	0.00001
5	1	100	100	0.0001
6	1	100	100	0.001
7	1	80	80	0.00001
8	1	80	80	0.0001
9	1	80	80	0.001
10	1	60	60	0.00001
11	1	60	60	0.0001
12	1	60	60	0.001
13	1	40	40	0.00001
14	1	40	40	0.0001
15	1	40	40	0.001
16	1	20	20	0.00001
17	1	20	20	0.0001
18	1	20	20	0.001
19	1	10	10	0.00001
20	1	10	10	0.0001
21	1	10	10	0.001
22	1	0	0	0.00001
23	1	0	0	0.0001
24	1	0	0	0.001

Table 1. The different cost scenarios for which the maximum operational performance per unit of time has been determined.

Work-in-process carrying costs include costs of capital needed for work-in-process, materials handling costs incurred for holding work-in-process, and product obsolescence and engineering changing costs. Depending on product type, this value may vary over a wide range. For instance, for a difficult to handle product with very short product life cycle a very high value might be appropriate. For other types of products lower values might apply. In order to capture this wide range of possible values, we have selected yearly work-in-process carrying costs equal to about 800%, 80% and 8% of the product value. Assuming a working year of 8000 hours, the work-in-process carrying costs on an hourly basis are r=0.001, r=0.0001, and r=0.00001, respectively.

The results of the optimization are given in the tables 2 and 3. The scenarios 1 to 3 and 22 to 24 show the results where switching the arrival rate is extremely costly resp. where switching does not incur any costs. In the next section we discuss these results.

Scenario	P	W_{lh}	W_{hl}	λ_h	λ_l	T	W	%high	F
1	0.958	1	600	0.96	0.96	0.960	120	99.84	$8.6\ 10^{-10}$
2	0.947	1	600	0.96	0.96	0.960	120	99.84	$8.6\ 10^{-10}$
3	0.864	18	600	0.93	0.44	0.930	66	100	0
4	0.982	328	600	0.99	0.97	0.987	359	86.60	$2.2\ 10^{-5}$
5	0.956	200	568	0.98	0.10	0.979	232	99.92	$3.8\ 10^{-6}$
6	0.864	64	255	0.93	0.10	0.930	66	100	$3.0\ 10^{-7}$
7	0.982	349	600	0.99	0.97	0.987	362	87.20	$2.2\ 10^{-5}$
8	0.956	202	541	0.98	0.10	0.979	229	99.90	$5.2\ 10^{-6}$
9	0.864	64	238	0.93	0.10	0.930	66	100	$7.6\ 10^{-7}$
10	0.983	293	600	100	0.97	0.990	420	65.95	$4.8\ 10^{-5}$
11	0.957	205	511	0.98	0.10	0.979	224	99.87	$7.6\ 10^{-6}$
12	0.864	65	218	0.93	0.10	0.930	66	99.80	$2.2\ 10^{-6}$
13	0.984	329	600	1.00	0.97	0.990	432	67.24	$5.2\ 10^{-5}$
14	0.956	209	475	0.98	0.10	0.978	218	99.83	$1.2\ 10^{-5}$
15	0.864	65	195	0.93	0.10	0.930	66	99.95	$7.2\ 10^{-6}$
16	0.985	385	600	1.00	0.97	0.991	453	69.02	$6.4\ 10^{-5}$
17	0.957	219	427	0.98	0.1	0.978	209	99.75	$2.2\ 10^{-5}$
18	0.864	66	165	0.93	0.10	0.929	64	99.82	$3.2\ 10^{-5}$
19	0.986	392	600	1.01	0.97	0.992	490	54.08	$1.0\ 10^{-4}$
20	0.957	172	342	0.99	0.92	0.980	215	85.45	$1.1\ 10^{-4}$
21	0.865	61	138	0.94	0.10	0.933	66	99.17	$1.8\ 10^{-4}$
22	0.987	599	600	2.0	0.93	0.994	614	5.940	0.120
23	0.960	196	197	2.0	0.10	0.980	196	46.32	0.945
24	0.878	59	60	2.0	0.10	0.937	59	44.03	0.936

Table 2. The operational performance (P) per unit of time and the corresponding optimal values for the switching points W_{hl}, W_{lh}, and the arrival rates λ_h, λ_ℓ, as found by OptQuest®, for the economic scenarios as given in Table 1 and work-in-process independent order processing rates. T = average throughput per unit of time; W = average work-in-process; %high = percentage of time that λ_h is used; f=switching frequency (average number of switches per unit time)

6. Discussion of results.

Work-in-process independent order processing rates.

Comparing the results for the scenarios 1 to 3 with the results for the scenarios 22 to 24 in Table 2, gives an estimate of the performance improvement that can be obtained from being able to influence the order arrival rate. This improvement ranges from 0.029 (about 3%) for low costs of carrying to 0.014 (about 1.6%) for high carrying costs of work-in-process. This seems to be a small improvement, but keeping in mind that profit margins in production systems are in the range of 10% of the product cost price, a 3% improvement in operational performance can increase the yearly profit with up to 30%. These improvements are obtained under zero switching

Scenario	P	W_{lh}	W_{hl}	λ_h	λ_l	T	W	%high	f
1	0.747	39	600	0.75	0.34	0.750	20	100	$3.0\ 10^{-9}$
2	0.745	39	600	0.75	0.10	0.750	19.8	100	$3.0\ 10^{-9}$
3	0.727	39	585	0.75	0.10	0.750	19.8	100	$3.0\ 10^{-9}$
4	0.835	46	120	0.85	0.32	0.846	37	99.24	$1.1\ 10^{-4}$
5	0.832	46	120	0.85	0.11	0.846	37.2	99.46	$1.1\ 10^{-4}$
6	0.798	42	113	0.85	0.10	0.846	36.6	99.47	$1.1\ 10^{-4}$
7	0.838	43	115	0.86	0.47	0.853	40	98.26	$1.8\ 10^{-4}$
8	0.835	43	115	0.86	0.38	0.853	40.0	98.60	$1.8\ 10^{-4}$
9	0.801	43	110	0.85	0.10	0.846	36.4	99.49	$1.0\ 10^{-4}$
10	0.842	46	113	0.86	0.41	0.854	40	98.57	$1.9\ 10^{-4}$
11	0.838	46	113	0.86	0.27	0.854	39.8	98.92	$1.9\ 10^{-4}$
12	0.803	41	105	0.86	0.10	0.854	38.7	99.16	$2.0\ 10^{-4}$
13	0.848	45	107	0.87	0.49	0.860	43	97.44	$3.0\ 10^{-4}$
14	0.844	45	106	0.87	0.45	0.860	42.3	97.71	$3.0\ 10^{-4}$
15	0.807	44	101	0.86	0.10	0.854	38.4	99.21	$2.2\ 10^{-4}$
16	0.858	44	96	0.89	0.59	0.871	47	93.77	$6.6\ 10^{-4}$
17	0.853	44	96	0.89	0.56	0.871	46.4	94.36	$6.8\ 10^{-4}$
18	0.814	43	91	0.88	0.10	0.867	41.8	98.31	$5.4\ 10^{-4}$
19	0.867	42	86	0.92	0.66	0.882	50	85.39	$1.5\ 10^{-3}$
20	0.863	43	87	0.91	0.62	0.880	48.6	89.56	$1.2\ 10^{-3}$
21	0.821	42	82	0.90	0.33	0.877	43.8	95.87	$1.2\ 10^{-3}$
22	0.916	46	47	2.0	0.10	0.916	46	42.96	0.932
23	0.912	45	46	2.0	0.10	0.916	45.3	42.96	0.932
24	0.871	45	46	2.0	0.10	0.916	45.3	42.95	0.932

Table 3. The operational performance (P) per unit of time and the corresponding optimal values for the switching points W_{hl}, W_{lh}, and the arrival rates λ_h, λ_ℓ, as found by OptQuest®, for the economic scenarios as given in Table 1 and work-in-process dependent order processing rates. T = average throughput per unit of time; W = average work-in-process; %high = percentage of time that λ_h is used; f=switching frequency (average number of switches per unit time)

costs. For positive switching costs a large part of this maximum improvement is maintained if the work-in-process carrying costs are small (0.00001) or medium (r=0.0001). However, for high (r=0.001) work-in-process carrying costs an insignificant performance improvement remains, even for switching costs as low as 10 (equal to about 0.5 –0.6 % of the yearly turnover). These results suggest that for production systems with work-in-process independent processing rates, order arrival rate control can be economically beneficial if work-in-process carrying costs are not too high. Order arrival rate control allows operating the system at a very high order arrival rate, resulting in high work-in-process and a high output rate, without running the risk that the work-in-process runs out of control. If work-in-process carrying costs are extremely high this option is costly and not used.

It is remarkable how infrequently that the arrival rate switching is used. Even for switching costs as low as 5% of the yearly turnover (the scenarios 19, 20 and 21) switching on average occurs about every two to two and half years (assuming 2000 working hours per year). For higher switching costs he average switching frequency is even smaller (note that with zero switching costs, about every hour the arrival rate is changed, resulting in about a constant work-in-process). The advantage of being able to change the arrival rate apparently is that the system can run at a rather high arrival rate for most part of the time, without having the risk that, due to the stochasticity of the arrival process, the work-in-process takes on extremely high values. These extreme values can be avoided by (infrequently) switching to a low arrival rate during some time.

Work-in-process dependent order processing rates.

Comparing the results for the scenarios 1 to 3 with the results for the scenarios 22 to 24 in Table 3, reveals the maximum performance improvement that can be obtained from being able to influence the order arrival rate in this type of production system. This improvement ranges from 0.169 (about 22.6 %) for high-work-in-process carrying costs to 0.144 (about 19.8%) for low work-in-process carrying costs. For positive switching costs much of the improvement is maintained for each of the three different work-in-process carrying cost scenarios. For instance with switching costs equal to 10 (about 0.6% of the yearly turnover) the performance improvement over the no-switching case is equal to 0.12 (about 16.1%) for low work-in-process carrying costs, to 0.118 (about 15.8%) for medium work-in-process carrying costs and to 0.94 (about 12.9%) for high work-in-process carrying costs. Even if switching costs are 10 times as high (representing about 6% of the yearly turnover) an 11.8% performance improvement can be achieved for low work-in-process carrying costs, and a 9.8% performance improvement can be achieved for high work-in-process carrying costs.

With work-in-process dependent order processing rates, arrival rate switching occurs more frequently. For instance, for the scenarios 19, 20 and 21 we see that switching occurs on average every 3 to 4 months (assuming 2000 working hours a year). For higher switching costs, switching on average occurs between about once per year (scenario 18) and once per 5 years (scenario 4).

The analysis of the data in the Tables 3 and 4 suggests that the use of arrival rate control can considerably improve the operational performance of production systems, in particular for production systems where the order-processing rate depends on the work-in-process. Operational performance improvements up to 16% are possible (which possibly may double the profit margin). The performance improvement seems to depend on the work-in-process carrying costs. High work-in-process carrying costs itself forces the system to operate at a low level of work-in-process, thereby to some extent sacrificing the higher throughput that can be realized at a higher level of work-in-process.

If order processing rates are independent of work-in-process, the performance improvement that can be obtained is much smaller, but still economically interesting (up to 3%). Here for systems with high work-in-process carrying costs no improvement seems to be possible. Table 2 shows that for low and medium work-in-process carrying costs the performance improvement is obtained by letting the system operate at a rather high average level of work-in-process, leading to a high throughput, while avoiding extremely high work-in-process levels by every now and then reducing the arrival rate.

7 Conclusions.

In this paper we investigated the operational performance improvement that can be obtained from being able to control the order arrival rate at a switching cost, in job shop-like make-to-order production systems. We investigated production systems where the order-processing rate is independent of the work-in-process and production systems where the order-processing rate is dependent on the work-in-process. We considered revenues per order, work-in-process carrying costs and order arrival rate switching costs. We have developed a model that allows the calculation of the operational performance as a function of the arrival rate switching parameters, and have used OptQuest® to determine the optimal switching parameter values for a range of economical conditions (work-in-process carrying costs and switching costs) for the two production systems considered. The results indicate that with work-in-process independent order processing rates and low to medium work-in-process carrying costs a small but economically relevant improvement in operational performance can be obtained. With work-in-process

dependent order processing rates, as modeled in this research, the use of arrival rate control can substantially improve the operational performance. Depending on the work-in-process carrying costs, improvements in operational performance between 9% and 16% can be obtained. These improvements are quite insensitive to the switching costs in a range of 0.6% to 6% of yearly turnover per switch in arrival rate.

In this research we have modeled switching costs as being independent of the size of the change in arrival rate. In some situations this may be unrealistic. In future research we will therefore investigate the situation where arrival rate switching costs consists of fixed costs and a cost that depends on the change in the arrival rate.

References

Bar-Lev, S.K., Parlar, M. and Perry, D. (1996), 'Analysis of a two-sided production-inventory policy with inventory-level-dependent production rates', *Applied Stochastic Models and Data Analysis,* Vol. 12, pp. 221-237.

Buss, A.H., Lawrence, S.H. and Kropp, D.H. (1994), 'Volume and capacity interaction in facility design', *IIE Transactions,* Vol. 26, Number 4, July 1994, pp. 36-49.

Cohen, J.W. (1976), 'On the optimal switching level for an M/G/1 queueing system', *Stochastic processes and their applications,* 4, pp. 297-316.

Doshi, B.T., van der Duyn Schouten, F.A. and Talman, A.J.J. (1978), 'A production-inventory control model with a mixture of back-orders and lost-sales', *Management Science,* Vol. 24, No. 10, pp. 1078-1086.

Hillier, F.S. (1963), 'Economic models for industrial waiting line problems', *Management Science,* Vol. 10, No.1, October, pp. 119-130.

Holström, J. (1994), 'The relationship between speed and productivity in industry networks: A study of industrial statistics', *International Journal of Production Economics,* vol. 34, pp.91-97.

Hopp, W.J. and Spearman, M.L., (2000), *Factory Physics*, Irwin McGraw-Hill.

Lieberman, M.B. and Demeester L. (1999), 'Inventory Reduction and Productivity Growth: Linkages in the Japanese Automotive Industry', *Management Science,* vol. 45, no. 4, pp.466-485.

Nishimura, A. and Jiang, Y. (1995), 'An M/G/1 vacation model with two service modes', *Probability in the Engineering and Informational Sciences,* 9, pp. 355-374.

Nobel, R. (1998), *Hysteretic and Heuristic Control of Queueing Systems,* Ph.D. dissertation, Amsterdam.

Schmenner, R.W. (1988), 'The Merit of Making Things Fast', *Sloan Management Review,* Fall 1988, pp. 11-17.

So, K.C. and Song, J. (1998), 'Price, delivery guarantees and capacity selection', *European Journal of Operational Research,* 111, pp. 28-49.

Stidham, S., JR. (1985), 'Optimal control of admission to a queueing system', *IEEE Transactions on Automatic Control,* Vol. AC-30, No. 8, August, pp. 705-713.

Tijms, H.C. (1977), 'On a switch-over policy for controlling the workload in a queueing system with two constant service rated and fixed switch-over costs', *Zeitschrift für Operations Research,* Band 21, pp. 19-32.

Tijms, H.C. and van der Duyn Schouten, F.A. (1978), 'Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate', *Stochastic Processes and their Applications,* Vol. 6, pp. 213-222.

Wickens, C.D., (1992), *Engineering Psychology and Human Performance*, Harper.

Xu, S.H. and Shantikumar, J.G. (1993), 'Optimal expulsion control-a dual approach to admission control of an ordered-entry system', *Operations Research,* Vol. 41, No. 6, pp. 1137-1152.