VARIABILITY IN ASSEMBLY AND COMPETING SYSTEMS: EFFECT ON PERFORMANCE AND RECOVERY

ROHIT BHATNAGAR AND PANKAJ CHANDRA
Faculty of Management and G.E.R.4.D, McGi11 University, Montreal, PQ H3A IG5 Canada

We examine the effect of variability on the performance of an assembly system where two stations feed another station and a competing system where two independent stations are fed by a single station. We study the impact of variability in processing times, variability due to unreliable workstations, and variability due to imperfect yield. The above impact is examined for both PUSH and PULL manufacturing environments. We also discuss design features of the two configurations that could be used to recover part of the output lost due to variability.

Manufacturing systems operate with different levels and diverse sources of variability in the production environment. The throughput of a manufacturing system has been shown to be closely related to the level and kind of variability present in the system (Conway *et al.*, [12]). Production systems should therefore be designed to manage variability by incorporating their effect in production and inventory planning. Variability could be introduced in production systems through unequal processing times random breakdowns of workstations, yield losses etc. Consequently the throughput of the system may be affected as some production capacity is lost due to variability. Over the long run management may be looking to improve equipment and processes in order to reduce (or even eliminate) variability . In the short run, some of this lost capacity can be re-covered by using work-in-process (WIP) inventory.

For a facility to operate at its maximum capacity and with no WIP, successive workstations that comprise the facility must be perfectly synchronized with regard to the instant at which they complete one part and start the next. i.e. , there should be no' 'blocking' .or 'starving' of workstations. Such synchronization is rarely achieved in real settings due to variability in the system. An important consequence of starving and blocking is that workstations have to remain idle which reduces the productive capacity of the facility. Hence, this problem has interested both practitioners and academic researchers. Use of adequate inter-stage buffers in a balanced line reduces this effect by decoupling work- stations. Effective management of WIP to reduce variability in the production environment requires determination of not only the quantity of buffer stock in the production system but also its location. This requires estimating the impact of variability on different configurations, e.g.. serial lines, assembly systems, etc. On the other hand, by carefully selecting processing rates, a system's throughput can be affected (Cumings, [15]). This is another line of inquiry that needs to be pursued, in conjunction with the role of buffers, to improve the throughput of a 'variable' system.

This research is aimed at understanding the impact of variability on the performance of two kinds of production systems. assembly and competing. An assembly system is one where parts from more than one station are assembled together to form a single part (Figure la). A competing system is one where more than one station competes for the output of a single station (Figure lb). We distinguish these systems from a serial line where workstations are connected in tandem (Figure 1 c) .The objective of this paper is to study the performance of assembly and competing systems in the presence of variability, to identify parameters of variability that affect performance, and finally to identify ways to recover part of the lost capacity .We study assembly systems for both PUSH and PULL manufacturing environments. The intuition gained from these results could be used by an operations manager to design production systems that can recover capacity lost due to variability by placing buffers of minimal size at appropriate locations. It also allows the practitioner to gauge the benefits obtained by using buffers which could then be compared with its costs to make an informed decision.

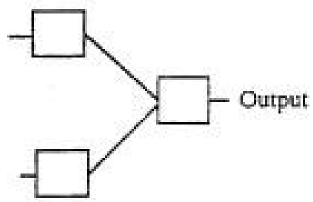

Figure 1a. Assembly system

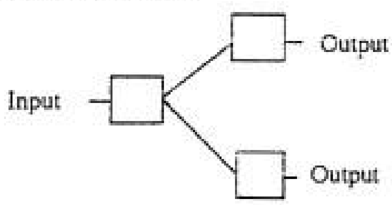

Figure 1b. Competing system

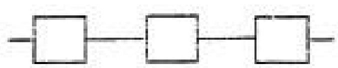

Figure 1c. Serial line

There is a substantial body of literature pertaining to the analytical study of serial lines (see Awate and Sastry, [4] and Dallery and Gershwin, [16] for a comprehensive review). These models, however, treat only the simplest configurations like the two-station serial line. For larger problems or even for a two-station line with restrictions, the above analysis becomes complex and yields intractable models (Commault and Dallery, [11]) All these models have focused on finding the throughput for serial lines" with pre-defined buffer capacities. However, while designing production facilities, managers are often confronted with a somewhat different problem. Given a global storage constraint, there is the question of determining optimal storage capacity that should be allocated to each buffer location; This important issue of optimally allocating storage capacities has been addressed by very few authors (see Altiok and Stidham, [2]). Moreover, all the above papers have focused exclusively on serial lines and do not treat more general structures often found in practice, e.g., assembly and competing systems.

There are few papers in the literature that study assembly and competing systems. Some of these have considered production facilities as queuing networks and have derived steady state performance measures. They assume infinite buffer space between successive work- stations which allows the analysis of each queue in isolation once the arrival process from other queues is determined. This assumption will almost certainly be violated in most real production systems. Assembly systems with infinite buffer space are inherently unstable even if the arrival processes are slower than the service process (Harrison, [20]). This would lead to infinite WIP in the system. Gershwin [18] shows that even with finite buffer space, very quickly the analysis of these models becomes intractable. He shows that a 20 ma- chine line with 19 buffer locations (of 10 units each) has over 6.41 X 10^{25} states in a Markov chain representation of the line. Most published work has, therefore, focused on approximate decomposition algorithms. The reader is referred to assembly-like queuing models of Bhat [6], BOI1Omi [8], Hopp and Simon [21], and Lipper and Sengupta [27] .Hopp and Simon show the equivalence of a three station serial line with buffers to a buffered 3-station assembly system under restrictive assumptions. Recent work of Gershwin [19] and Di Mascola *et ai.* [17] on systems liable to breakdown t'al1 in this category of decomposition algorithms where assembly/disassembly systems are decomposed into several two-station serial lines with buffers. As Gershwin [19] points out, the inherent problem with the above approaches is that assembly systems are not decomposable exactly. Moreover, the "evaluation sequence which specifies the sequence in which the evaluation of sub-systems proceeds may fail to give a convergent solution. The procedure of Di Mascola *et al.* [19] is more reliable but extremely complicated thereby limiting its use by managers. Alves' [3] study on the performance of serial/parallel lines uses a Markov chain approach. Consequently, it suffers from the problems discussed above. The assembly system considered here needs to be distinguished from a "merge" configuration often used in queuing models. The physical assembly of multiple parts into a single part leads to the consumption and consequently the effective disappearance of the multiple parts from the process. A "merge" configuration, on the other hand, continues to retain the identities of the multiple parts or customers in the queue. This makes such a queuing analysis of the assembly system very complicated and often inaccurate. For a more general review of modeling manufacturing systems the reader is referred to Buzacott [10].

None of the above models for assembly systems have addressed the problem of determining optimal allocation of global buffer capacity among different locations or given insights into their design. Given the complexity of the analysis involved, this seems to be an extremely difficult problem to approach analytically. At the same time, the best results from analytical models for this problem in serial lines seems to be no better than simulation results (Jafari and Shanthikumar, [23]). Another problem for which no analytical results are available for assembly systems is the case where workstations have variable processing rates. Gershwin [18] addresses this problem for serial lines but no analytical results or bounds are reported.

Finally, we note that there have been very few research attempts to analyze the perfonnance of competing systems. Such systems are often found in practice especially when using a backup machine for reducing variability due to unreliability or variable processing times. While analytical results for serial lines by Ignall and Silver [22] and Kubat and Sumita [25] are in this direction, no insights are provided for design of complicated competing systems. Lee and Pollock [26] model systems that are combinations of tandem split (or competing) and merge configurations while assuming exponential service times. Such models (including Perros and Snyder, [29] and Altiok and Perros. [1]) suffer from the disadvantage mentioned earlier for merge configurations. Once again, the issue of sizing and allocation of buffer space to various locations is not addressed in these models.

Given the limitations of these methods alternative tools need to be used. Simulation seems to be a good direction to take for two reasons. First, it helps us overcome the computational difficulties of Markov chains and the complexity and inaccuracies of decomposition techniques. Second, it is a good first step for providing vital operational- insights and intuition as to how assembly and competing systems should be designed (which is missing in the research listed above). Well crafted simulation can be used to identify critical parameters which will help in a more focused work on building efficient approximate algorithms.

Our line of inquiry in this research closely resembles two excellent earlier papers, Conway, Maxwell, McClain and Thomas [12] -henceforth CMMT and Baker, Powell and Pyke [5] henceforth BPP. In their pioneering work, CMMT provide insights into the role of buffers in serial lines. They show that interference between workstations in balanced serial lines results in a loss of capacity of the line but most of this loss occurs in the first few stations. This loss of capacity may be reduced by placing buffers between workstations. Increased buffer size, however, shows very rapidly decreasing rates of return. Recovery of lost output is found to be dependent as much on the actual positioning of the buffer as on its size. BPP look at effects of buffers on the throughput of assembly systems under PUSH and PULL modes of production. They show that throughput in a PUSH mode exceeds that in a PULL mode. BPP found that the mode of operation is critical in determining the best location of buffers in a 5 and 7-station assembly line. These results need to be extended for unreliability and yield related variability.

This paper is organized as follows: In the next section we explain the methodology used to study the impact of variability. Then we present the results for assembly systems operating under three sources of variability mentioned earlier. We also show how through- put changes when the processing rates of the workstations are varied. Next we outline the role of processing time variability in competing workstations. Finally, in the last section we summarize our findings and recommendations concerning the design and operation of the two configurations.

Methodology

As in CMMT and BPP, our study used the simulation package XCELL (Conway *et al.,* [14]) for studying the configurations under the PUSH mode of manufacturing while for the PULL mode we used XCELL + (Conway *el at.,* [13]). The simulation environment was quite similar to those mentioned in these two related efforts. The average processing time for each operation is taken as 1.0. Processing at the assembly station begins only when both inputs are available to it, irrespective of the mode of operation.

Throughput here is used to denote the steady state output per unit time. To study processing time variability we initially conducted simulation (or data collection) runs of 10,000 time units following a warm up period of 100 time units (similar to the experimental designs of CMMT & BPP). However, we applied two additional criteria to be more certain that the throughput figures represent steady state values. First, for each configuration we made five independent runs using different random seed numbers and the throughput values reported are averages over these runs. Second, some simulations were re-run with altered starting conditions. For all the simulation runs, we ensured that the maximum deviation in throughput with different starting condition or a different seed number did not exceed 1%. We found that for truncated normal and uniform distributions our criteria were satisfied by a warm-up period of 100 time units and a data collection period of 10,000 time units. However, with exponential processing times, the data collection period had to be increased to *50,000 time* units in order to reduce the deviations to below 1% .To study variability due to unreliability, a warm- up period of l0,000 time units and a data collection period of *300,000* time units were required to satisfy our criteria. Further details are provided in the following sections.

Assembly Systems

In this section we focus our attention on a 3-station assembly system (Figure la). Two components are first processed at independent stations and then assembled on the assembly station. For reference, we shall call the independent stations as “serial stations.” Under the PULL mode, the transfer of completed pans from the serial stations to the assembly station takes place only when both parts are ready. However, under the PUSH mode, the assembly station can accept a part from one of the serial stations while awaiting the other. We in- tend to illustrate how different types of variability in the manufacturing process affect the output and how can we employ WIP effectively to recover this lost capacity .As pointed out by BPP, the throughput varies depending on the mode of production. The impact of variability is found to be at least as much for a system in a PULL mode as compared to operating it in a PUSH mode. We illustrate his phenomenon for the three sources of variability under study. Our observations underline the importance of considering the mode of operation while designing production systems.

Variability in Processing Times

In order to isolate the effect of variability of processing times, we study 3-station assembly systems that are perfectly reliable and have 100% yield. Variability is induced by the workstation processing time distribution. CMMT [12] have studied serial lines under these conditions. When workstations have constant processing times (coefficient of variation, c. v. = 0.0) there is no interference between adjacent workstations as they start and [mish a part at the same instant. The output in this case represents the theoretical maximum output possible in such a system (i.e., normalized to a value of 1.0 here). However, when there is variability between processing times of adjacent workstations, output is reduced to a value below this theoretical maximum level. For serial lines, CMMT show empirically that the reduction in output is dependent on the c. v. of the work-station processing time distribution (see Muth. [28] for analytical results).

BPP [5] have extended these results of CMMT, again empirically, to assembly systems. They suggest that the reduction in output caused by processing time variability is dependent on the coefficient of variation. Bhatnagar and Chandra [7] illustrate these results for varying levels of c. v. These reductions in throughput result in a considerable loss of capacity in such assembly systems. Hence. it would be prudent for production managers to invest resources wherever possible to reduce this variability .

CMMT [12] have suggested that one way of recovering lost capacity is by providing for butters between workstations. The provision of buffers for storage of WIP between workstations can bring about a certain degree of decoupling of adjacent workstations by reducing idleness due to starving and blocking. Figure 2 shows a 3-station assembly system with buffers where Wl and W2 are the two serial stations and W3 is the assembly station. We study the effect of buffer size at locations B 1 and B2 (Figure 2) for three processing time distributions. namely uniform, exponential and truncated normal. We assume that the same buffer size is used at both locations, i.e., for buffer size 1, both locations have buffer of size 1. Let V_T be the throughput with O buffer for the operation mode T (T = PUSH or PULL). Then the throughput lost due to processing time variability is given by $(1 - V_T)$ since the theoretical maximum throughput (at c. v. = 0.0) in our case is 1.0. Our results show that under PUSH mode, the through- put lost is 0.080 in the truncated normal case (c. v. = 0.115), 0.181 in the uniform case (c. v. = 0.289) and 0.422 in the exponential case (c. v. = 1.00). For these distributions (with the same c. v. as above), the through- put lost under the PULL mode of operation is 0.091, 0.200, and 0.451 respectively (see Bhatnagar and Chandra [71 for details). Percent recovery of throughput is $\{(Vi - V_T)/(1 - V r)\} * 100$ where Vi is the throughput with buffer size at each position equal to i. The extent of lost capacity recovered by using buffers is shown in Table 1. For all three processing time distributions, the percentage recovery for a Buffer/c. v. ratio of 10 is in the range of75 to 85, which is the same as CMMT found for serial lines. The c. v. therefore seems to be the critical criterion for studying the impact of processing time variability since percentage recovery of an assembly system is dependent on the c. v. It should also be noted that with large buffer size, throughput will approach the theoretical maximum output of 1.0.

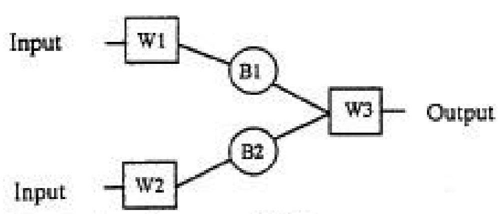

Figure 2. Assembly system with buffers

While buffering increases the throughput of the system and ensures the recovery of part of the lost output, it also imposes a cost on the system. This cost must be seen in terms of two distinct components: a *fixed cost* of setting up a buffer facility of requisite size and a *variable cost* which is incurred by carrying WIP in the system. The WIP is the total average work-in-process in the system. Figure 3 shows the interaction between the throughput and the two variables, buffer size and average WIP in PULL systems (for the same simulation run as Table I) for different types of processing time distribution. See Bhatnagar and Chandra [7] for a comparative figure of the system under the PUSH mode. The throughput vs. buffer plot indicates the buffer size necessary to obtain a desired output from the given system. For this buffer size the WIP vs. buffer plot gives the approximate value of the average WIP

Table 1. Percent recovery via Buffers for Various Processing Time Distributions

	Truncated Normal		Uniform		Exponential	
Buffer size	PUSH	PULL	PUSH	PULL	PUSH	PULL
0	0	0	0	0	0	0
1	80	82.4	59.7	63	22.7	26.6
2	88.3	90.1	75.1	77.5	37.4	41.2
3	92.5	93.4	81.8	83.5	47.6	51
4	93.8	94.5	86.2	87.5	54.3	57.2
5	95	95.6	88.4	89.5	59.7	62.1
6	95	95.6	90.6	91.5	64.9	66.3
7	96.3	96.7	91.7	92.5	67.5	69.6
8	96.3	96.7	92.3	93	70.1	72.1
9	97.5	96.7	92.8	93.5	73.5	74.9
10	97.6	97.8	93.9	94.5	75.4	76.7
1000	98.3	98.9	98.3	98	97.6	96.7

that will have to be carried in the system. The **throughput vs WIP** plot is the combination of the above plots and represents what Karmarkar [24] calls the "clearing function." It allows the operations manager to evaluate the accompanying costs due to the build up of WIP when a certain throughput level is desired. Consequently, the cost of increased throughput can be determined.

An alternative to buffering is to increase the processing rate of different workstations. Small increases in the processing rate may be possible on some machines. Larger increases may be seen as making additional work- stations available so as to form a type of parallel system. As BPP observe, the choice between buffering and increasing the processing rate is essentially an economic issue governed by the relative costs of the two alternatives. However, the designer of a production facility may find it useful to know the maximum increase in throughput possible by implementing each of these alternatives. BPP consider such an improvement by increasing the processing rate of the assembly station and conclude that the maximum throughput thus possible (i.e., when the assembly station operates at infinite speed) is less than the maximum throughput that can be obtained by buffering. This result may be violated under the PUSH mode of operation. Under the PUSH mode, the asymptotic improvement in throughput possible by increasing the processing rate of the assembly station is comparable to that obtained by increasing buffer size when the c .v. is low. One possible explanation is that the serial stations have a higher probability of finding the queue for the assembly station empty, which reduces the blocking of serial stations. We also note that under the PUSH mode, the impact of increasing the processing rate is more prominent for the assembly station relative to that of the serial stations. This is explained by the fact that the throughput achieved by increasing the processing rate of a serial station is bounded by the throughput of a 2-station serial line. There is no difference between the two under the PULL mode of operation as the system acts like a synchronous line.

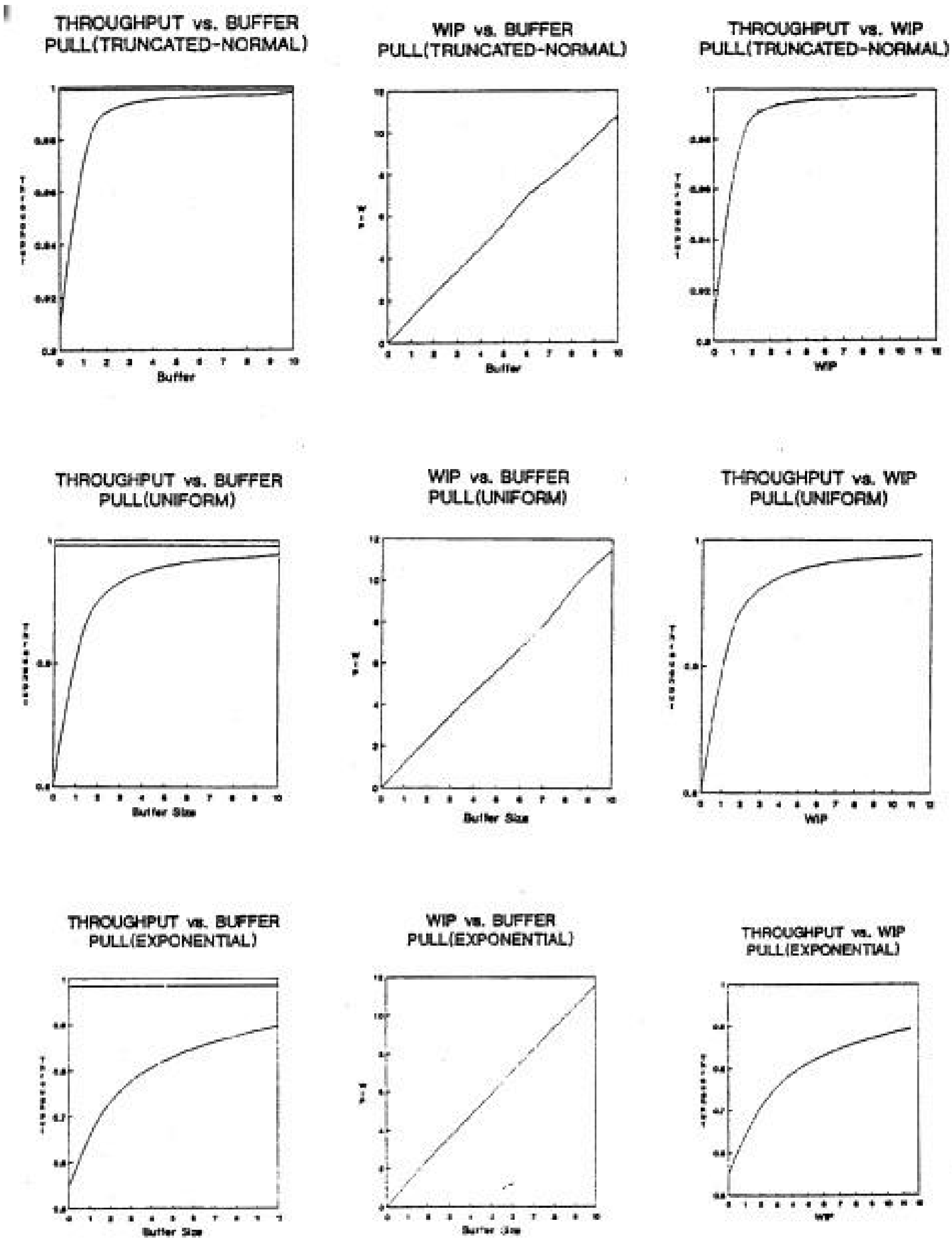

Figure 3. Performance and clearing functions for an assembly system with different processing time distributions

Theoretically, the assembly station is not the only workstation whose speed can be increased. It may in fact be possible to speed up any/all stations. We study the effect of increasing the processing rate of workstations on the output of a 3-station assembly system by studying the following cases: (i) increase processing rate of one serial workstation; (ii) increase processing rate of the assembly workstation; (iii) increase the processing rate of both serial workstations; (iv) increase the processing rate of one serial workstation and the assembly work-station; and (v) increase processing rate of all workstations (assembly and serial workstations) . The increase in processing rate is modeled by reducing the mean processing time (while keeping c. v. constant). By considering the assembly system as 1 whole, the maximum increase in throughput possible by increasing the processing rate of workstations is in fact larger than the maximum increase possible by the use of buffers (Figure 4). These results give the designer of manufacturing systems an accurate idea of the exact location at which buffers or increase in processing rate could provide most benefit.

Unreliable Workstations

In order to isolate the effect of unreliability, in this section workstations have no variability due to processing times or due to imperfect yield. As in earlier cases, the processing time is taken as 1. The following variables are used: *U* = Mean time to failure; *D* = Mean time to repair; and *R* = *U/D.* In XCELL the processing of the current job is interrupted at the instant at which the station fails and is resumed as soon as the station is repaired. Time between failures is exponentially distributed with mean *U* while *D* is taken as constant. The values *of R* (1,5, LO, 120) represent a wide range of systems, from very unreliable to very reliable, and would cover most systems found in real situations. For each value of *R,* three different values of *D* are used and the corresponding value of *U* is chosen to be *R*D.* For each pair of U and *D* values, we ran five different cases depending on which set of workstations are un- reliable: one serial workstation is unreliable; assembly workstation is unreliable; both serial workstations are unreliable; one serial workstation and the assembly workstation are unreliable; and all workstations are un- reliable. Initially, we consider systems without intermediate buffers.

A data collection period of 300,000 time units was used following a warm-up period of 10,000 time units. Simulation runs using the above design were executed for both PUSH and PULL modes of manufacturing. In order to be certain that the throughput figures actually represent steady state throughput we conducted runs with different seeds, conducted runs with different warm-up periods and compared data at every 100,000 time units. For the last two cases shown in Table 2, i.e., *U=3600, D=30* and *U=7200, D=60,* the data collection periods were increased to 500,0000 and 1,000,0000 time units respectively. In all cases we found that maximum deviation in throughput \vas less than 1 %. This confirmed our belief that the sample size was adequate for the estimation of steady state throughput rates.

The results of various runs are summarized in Table 2. For a given value of *R,* we found that throughput is relatively insensitive to the individual values of *U* and D. *R* appears to be the critical factor, which accounts for the reduction in output due to unreliability of work- stations. We also observe that throughput varies only with the number of workstations that are unreliable irrespective of whether they are assembly or serial work- stations. That is, throughput is the same if anyone of the workstations Wl, W2 or W3 is unreliable or similarly if any two of the workstations are unreliable.

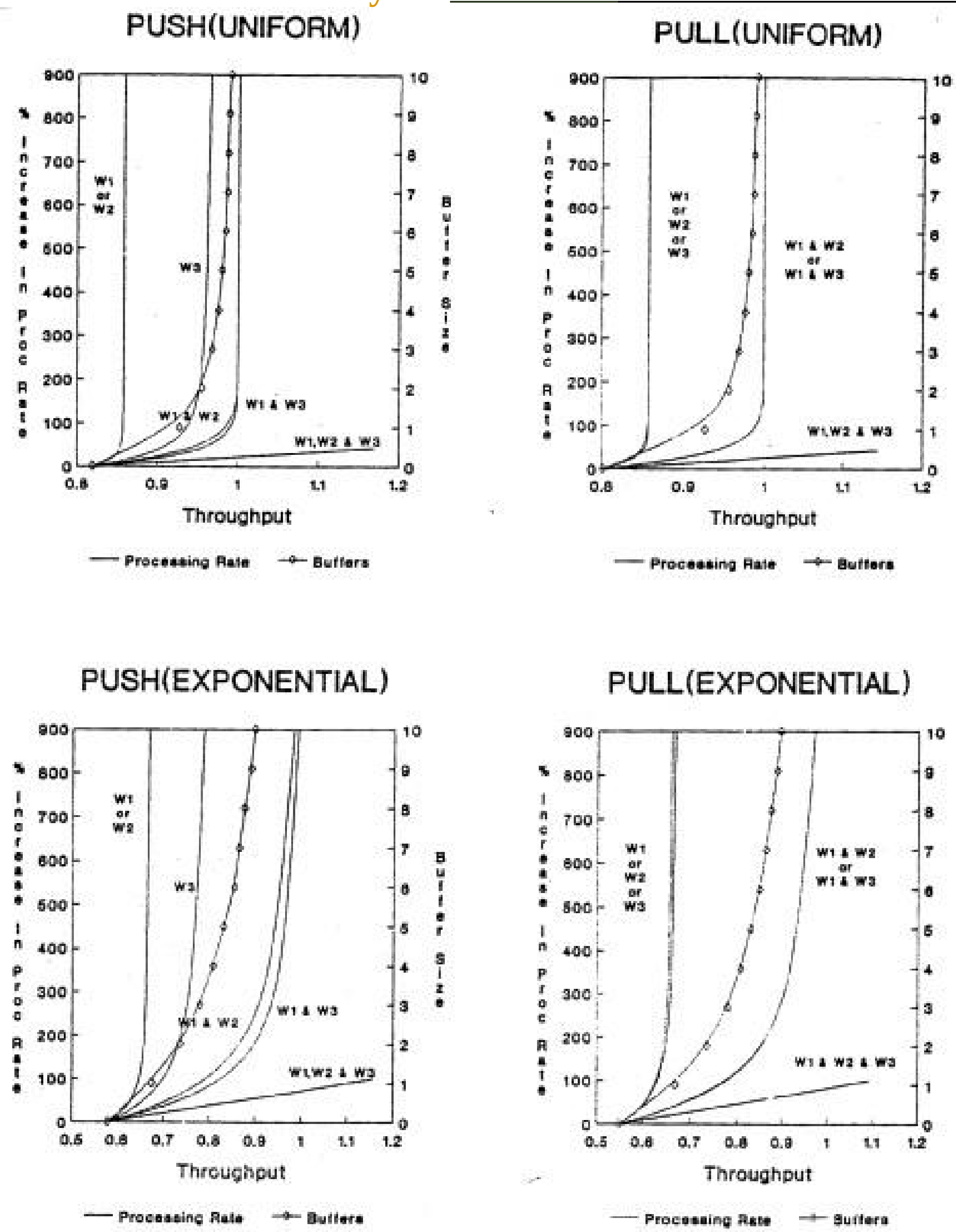

Figure 4. Comparison of recovery through buffers and increase in workstation processing rates

Table 2. Throughput of Unreliable Workstations

MODE	R	U	D	Unreliable Workstations: W1 or W2 or W3	W2 & W3	W1, W2 & W3
PUSH	1	5	5	0.525	0.378	0.32
		30	30	0.506	0.342	0.263
		60	60	0.502	0.34	0.257
	5	25	5	0.836	0.723	0.643
		150	30	0.832	0.715	0.629
		300	60	0.835	0.715	0.626
	10	50	5	0.91	0.836	0.777
		300	30	0.912	0.831	0.771
		600	60	0.91	0.835	0.771
	120	600	5	0.992	0.984	0.976
		3600	30	0.991	0.984	0.976
		7200	60	0.991	0.983	0.975
PULL	1	5	5	0.5	0.353	0.284
		30	30	0.499	0.336	0.256
		60	60	501	0.335	0.251
	5	25	5	0.833	0.716	0.635
		150	30	0.832	0.715	0.623
		300	60	0.834	0.714	0.626
	10	50	5	0.909	0.834	0.772
		300	30	0.909	0.831	0.77
		600	60	0.91	0.835	0.77
	120	600	5	0.992	0.984	0.976
		3600	30	0.991	0.984	0.976
		7200	60	0.99	0.983	0.975

Under the PULL mode, the throughput in the unbuffered assembly can be shown equal to $(U/(U + nD))$ where n is the number of unreliable workstations. The reader will recognize this expression as the same formula that Buzacott [9] had derived for synchronous serial lines (where the transfer to the next station is simultaneous for all workstations). This is intuitive because for synchronous lines without buffers, the throughput of a different 3-station assembly system is equivalent to that of 3-station serial line. Under the PUSH mode, the above formula yields only a lower bound (but is very tight as can be see from our results). This formula holds for small ratios of processing time to the mean time to failure and for large values of *R*, for which there is a high probability that at most one station is under repair at any instant. On the other hand, with small value of *R*, the formula holds for large values of *U*.

Buffers play a similar role in reducing variability due to unreliability as in the case of processing time variability. CMMT have proposed the idea of using buffers of size *D/P*, i.e. .the number of units a workstation could have produced when in repair. In assembly for a given *R*, the throughput achieved by using buffers of a given size would depend on the individual value of D (unlike the unbuffered case where for a given R throughput is dependent of the individual values of U and D). Table 3 summarizes the results of our simulation runs with buffers. For the purpose of illustration we give results for the PULL mode. We note that for a given value of R, percentage recovery varies inversely with D. Our conjecture, therefore is that for different values of D. Our conjecture, therefore, is that for different values of D if buffers are used in proportion to D/P one would recover same percentage of lost capacity in all cases. We ran an experiment and the results confirm our belief. With R=5, using buffers of 5,30 and 60 in three different cases (Case 1: U/D=25/5;Cases 2: U/D=150/30; Case 3: U/D=300/60) percentage recovery was 30.4, 31.0 and 30.7 respectively. Table 3 also shows that with fixed D, percentage recovery improves for a given buffer size if U is increased. Use of buffers larger than D/P are useful as they reduce the starving of the assembly station when two or more consecutive breakdowns occur at the same station or at close to the same time. The variability in the system increases when the downtimes are not deterministic. Therefore, it appears likely that larger buffers will be needed to achieve throughputs similar to those obtained with deterministic downtimes.

Imperfect Yield

Imperfect yield of good units by workstations is another source of variability that affects the throughput of a manufacturing system. Throughput is reduced since each time a workstation produces a bad unit the effective processing time of the next good unit is increased by the processing time of the bad unit.

We study the impact of imperfect yield in a 3-station assembly system under two cases-first when a serial workstation has poor yield and second when the assembly workstation exhibits poor yield. In each case the yield of other workstations is assumed to be perfect. We assume that rejects are scrapped. Six different reject levels are considered: 0% (i.e.. perfect yield), 10%, 20%, 25%, 50%, and 90%, where each unit is rejected with the specified probability .For each % reject level we look at three cases with different processing time distributions: Constant (c. v. = 0.0), uniform (c. v. = 0.3) and Exponential (c .v. = 1.0) .In each of these cases the percentage reduction = $\{(V_o - V_i)/V_o\}$ * 100, where V_o is the throughput with perfect yield while V_i is the throughput with % reject level *i*. Results are given in Table 4.
Where serial workstations Wl or W2, produce bad units the reduction in output due to imperfect yield (as compared to the case with perfect yield) is inversely related to c. v. of the processing time distribution. Thus, for the cases under study, the maximum reduction occurs for constant processing times while minimum reduction occurs for exponential processing times. This result is intuitive since in the case where processing times are constant the entire system remains idle for exactly the amount of time that was spent processing the unit which turned out to be bad and was rejected (1.0 in our case).

Table 3. Percentage Recovery with All Station Unreliable

	R = 5. U = 25, D = 5		R = 5. U = 150, D = 30		R = 5. U = 300, D = 60	
Butter Size	Throughout	%Recovery	Throughout	%Recovery	Throughout	%Recovery
0	0.635	0	0.623	0	0.626	0
1	0.663	7.7	0.626	0.8	0.627	0.3
2	0.689	14.8	0.632	2.4	0.629	0.8
4	0.728	14.8	0.634	5.3	0.635	2.4
8	0.765	35.6	0.662	10.3	0.645	5.1
1000	0.831	53.7	0.826	53.8	0.823	52.7
	R = 10. U = 50, D = 5		R = 10. U = 300, D = 30		R = 5. U = 150, D = 30	
Butter Size	Throughout	%Recovery	Throughout	%Recovery	Throughout	%Recovery
0	0.772	0	0.77	0	0.77	0
1	0.79	7.9	0.773	1.3	0.771	0.4
2	0.809	16.2	0.776	2.6	0.773	1.3
4	0.837	28.5	0.781	4.8	0.776	2.6
8	0.863	39.9	0.794	10.4	0.783	5.6
1000	0.908	59.6	0.904	58.8	0.904	58.3

Table 4. Imperfect Yield from Workstations

		% Reduction in Throughput			
		W1 or W2			W3
Mode	%Reject	Constant	Uniform	Exponential	All Distributions
PUSH	0	0	0	0	0
	10	10	6.5	4.2	10
	20	20	14.4	8.5	20
	25	25	18.3	11.8	25
	50	50	41.8	29.2	50
	90	90	87.8	83	90
PULL	0	0	0	0	0
	10	10	6.5	4.2	10
	20	20	13.9	8.4	20
	25	25	18.1	11.3	25
	50	50	41.4	29	50
	90	90	87.5	82.1	90

Hence, the reduction in output in this case is exactly equal to the reject percentage. However, if the c. v. is greater than 0.0 the configuration is not perfectly synchronized and part of the starving/blocking due to imperfect yield may occur at the same instant as the starving/blocking due to the processing time variability .As a result, poor yield leads to a higher percent reduction in the throughput of perfectly synchronized configurations than in those configurations where work- stations also have processing time variability.

In the case where the assembly workstation, W3, gives imperfect yield, the reduction in output is approximately equal to the reject % in all cases (Constant, Uniform and Exponential processing times). This case is equivalent to a 3-station assembly system followed by an inspection stage. Inspection is perfectly reliable and instantaneous. The throughput of this system will be defined by the product of the throughput of the 3-station assembly system (for a given processing time distribution) and the yield percentage. Our results confirm this.

Competing Systems

In this section we study the effect of processing time variability on throughput of systems where two work- stations are competing for the output of a single work- station. Four different configurations are shown in Figure Sa. Configuration I is the pure competing sys- tem. The other configurations are combinations of the serial line, assembly system and the competing system.

The average processing time for a part at each work- station is taken as 1.0. For each of the configurations, processing time variability was studied using uniform (c.v. = 0.289) and exponential (c.v. = 1.0) distributions.
In Configuration I, the throughput 5 (throughput 5 = throughput 5 1 + throughput 52) will depend on the extent to which W 1 is blocked. This can occur when W2 and W3 are busy at the instant when Wl completes processing a part. However, the probability of this event is low unless the workstations have very high processing time variability. With uniform processing times the throughput is 1.0 while it is 0.908 in the exponential case. Clearly in the first case (i.e., low c.v.) Wl was never blocked while in the second case (i.e., high c.v.) some blocking of Wl occurred. Consequently, addition of buffers at B (Figure 5b) is worthless in the first case while it can recover some lost capacity in the second case (where with *B=5,* 5=0.989 and average WIP=0.300; see Bhatnagar and Chandra [7] for details).

Configuration II combines assembly and competing systems. W2 is once again the bottleneck workstation and one would normally expect a throughput similar to the corresponding case in Configuration I. However a reduction in throughput may occur due to the presence of additional workstations W 1 and W3 which also pro- vide input to the downstream workstations W4 and W5. This degradation of throughput could take place as follows: (a) at the instant when W2 completes a part both W4 and W5 are idle and are awaiting input from their respective upstream workstations; or (b) at the instant when W2 completes a part one of the workstations W4

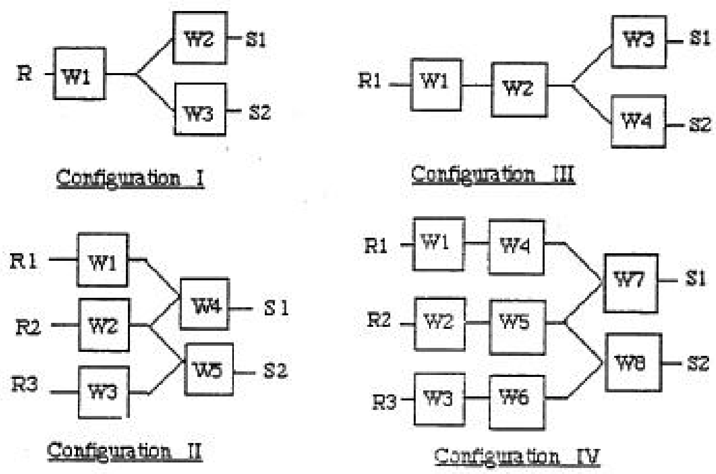

Figure 5a. Competing systems

or W5 is busy while the other is idle and awaiting input from its upstream workstation. In both cases W2 will be blocked with the frequency of blocking depending on the c. v. of the processing time distribution. The throughput figures of 0.995 and 0.821 in Table 5a validate this view. Table 5a gives the throughput when equal buffers of the size shown are placed at each of the three possible locations Bl. B2 and B3 (Figure 5b).

The relative WIP levels in buffers B 1. B2 & B3 for the exponential case in Table 5a motivated us to consider the issue of best allocation of a fixed number of buffer slots among these three candidate positions. That is, if a limited number of buffer slots were to be made available where would it be most beneficial to allocate them? Our results indicate that B2 is the best position, at least for a small number of buffer positions (Table 5b) .The optimal allocation for this type of configuration also exhibits BPP's ' buildup property,' i.e., the optimal allocation for $(n + 1)$ buffer slots builds up from the optimal allocation for n slots.

Configuration ill is a combination of serial and competing systems. The blocking of W2 in this case will be similar to Configuration I. However, W2 will additionally be subject to 'starving', since it is part of the serial line W 1 -W2 .We therefore expect lower throughput in this case as compared to Configuration I. This is confirmed by our simulation results (Table 6). Ll and L2 (Figure 5b) are the two possible positions of buffers. With large buffers at Ll. the upper bound on throughput S is the throughput of a 2-station serial line (0.85 and 0.65 with a c. v. of 0.289 and 1.0 respectively). While with large buffers at L2 the upper bound is the throughput of Configuration I.

Configuration IV is a combination of serial, assembly and competing systems. In this case we also have two possible locations for buffers-LI and L2 .In our experimentation, for each location (LI or L2) the same size buffer is placed on each of the three serial lines. The results are similar to those obtained for Configuration ill. With infinite buffer at L2 the upper bound on throughput is given by the unbuffered throughput of Configuration II while with infinite buffer at L 1 the upper bound on throughput is given by the throughput of an unbuffered 2-station serial line. Since the throughput is greater with buffer at L2 than at Ll, L2 should be the preferred location. Table 7 summarizes our results for this configuration. (For each location, BI, B2 and B3 ~correspond to buffers on individual input streams Rl, R2 and R3 respectively).

Two practical insights arise out of our study of these configurations for locating buffers in such manufacturing scenarios. First, in a configuration comprising a serial line along with competing workstations (with or without assembly stations) it is always preferable to place buffers in such a way so as to minimize the starving of the station which feeds all the competing workstations (e.g. W2 in Configuration ill and W5 in Configuration IV), i.e., place buffer upstream of such a workstation. Second, in a configuration without a serial line, e.g., competing workstations with assembly systems, if a limited number of buffer slots are available they should be placed immediately downstream of the workstation, which feeds all the competing workstations (e.g., W2 in Configuration 11).

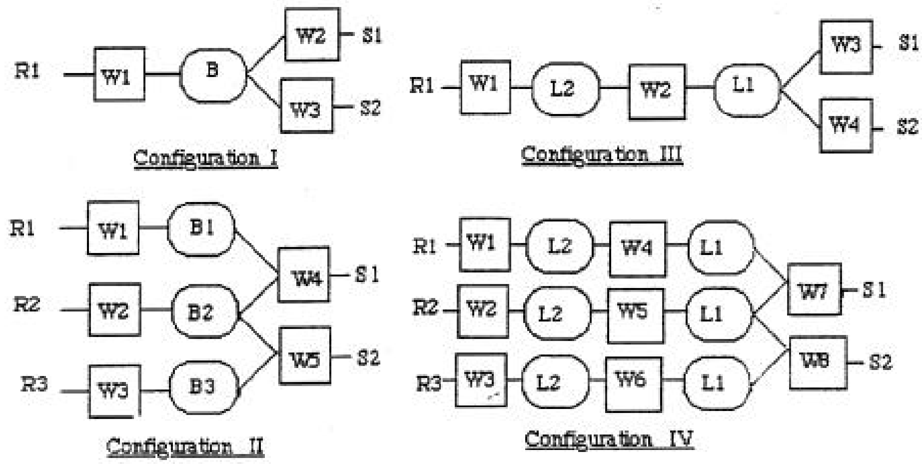

Figure 5b. Competing systems with buffer locations

Table 5a. Variability and Recovery in Competing Workstations-Configuration II (T= Throughput)

Buffer at B1, B2 & B3	Uniform T	Average WIP B1	B2	B3	T	Average WIP B1	B2	B3
0	0.995	0	0	0	0.821	0	0	0
1	0.999	1	0	0.96	0.935	0.86	0.19	0.8
2	1	2	0	1.96	0.964	1.78	0.24	1.67
3	1	3	0	2.95	0.979	2.75	0.3	2.58
4	1	4	0	3.96	0.988	3.71	0.35	3.44
5	1	5	0	4.96	0.999	4.7	0.33	4.45

Table 5b. Buffer Positions and Performance for Configuration II

	Allocation				WIP		
Buffer Slots	**B1**	**B2**	**B3**	**Throughput**	**B1**	**B2**	**B3**
0	0	0	0	0.821	0	0	0
1	0	1	0	0.89	0	0.28	0
	1	0	0	0.857	0.87	0	0
	0	0	1	0.856	0	0	0.81
2	0	2	0	0.923	0	0.52	0
	1	1	0	0.923	0.82	0.24	0
	2	0	0	0.864	1.8	0	0
3	0	3	0	0.951	0	0.72	0
	1	2	0	0.945	0.82	0.4	0
	2	1	0	0.923	1.73	0.22	0
	3	0	0	0.862	2.77	0	0
4	0	4	0	0.967	0	0.95	0
	1	2	1	0.965	0.84	0.34	0.77
	1	3	0	0.964	0.8	0.57	0
	3	1	0	0.923	2.68	0.2	0
	4	0	0	0.874	3.72	0	0
5	0	5	0	0.978	0	1.08	0
	1	3	1	0.977	0.83	0.44	0.77
	1	4	0	0.976	0.78	0.68	0
	2	2	1	0.967	0.86	0.3	1.64
	2	1	2	0.943	1.8	0.16	1.67

Table 6. Variability and Recovery in Competing Workstations-Configuration

	Buffer at L2				Buffer at L1			
	Uniform		Exponential		Uniform		Exponential	
Buffer Size	Throughput S	Average WIP	Throughput S	Average WIP	Throughput S	Average WIP	Throughput S	Average WIP
0	0.856	0	0.655	0	0.856	0	0.657	0
1	0.946	0.5	0.734	0.54	0.856	0	0.664	0.04
2	0.968	0.97	0.777	1.1	0.856	0	0.666	0.05
3	0.976	1.48	0.806	1.67	0.856	0	0.667	0.05
4	0.981	1.97	0.826	2.32	0.856	0	0.667	0.05
5	0.985	2.58	0.845	2.89	0.856		0.667	0.05
100	0.993	43.08	0.906	87.04	0.856	0	0.667	0.05

Table 7. Variability and Recovery in Competing Workstations - Configuration IV (T = Throughput)

Buffer Location	Buffer Size	Uniform				Uniform			
			Average WIP				Average WIP		
		T	B1	B2	B3	T	B1	B2	B3
L2	0	0.858	0	0	0	0.638	0	0	0
	1	0.946	1	0.52	0.99	0.708	0.94	0.58	0.9
	2	0.964	2	1.03	1.99	0.741	1.91	1.24	1.86
	3	0.975	3	1.59	2.99	0.768	2.9	1.88	2.84
	4	0.979	4	2.02	3.99	0.785	3.88	2.62	3.82
	5	0.984	5	2.74	4.99	0.789	4.89	3.51	4.8
	100	0.993	99.91	21.82	99.54	0.823	99.74	95.93	99.59
L1	0	0.855	0	0	0	0.637	0	0	0
	1	0.855	1	0	0.96	0.661	0.94	0.06	0.83
	2	0.856	2	0	1.96	0.667	1.91	0.05	1.71
	3	0.856	3	0	2.96	0.667	2.93	0.06	2.63
	4	0.856	4	0	3.96	0.667	3.93	0.05	3.63
	5	0.856	5	0	4.96	0.667	4.91	0.05	4.59
	100	0.856	99.86	0	98.83	0.667	99.92	0.05	99.59

Conclusions

The goal of this study was to understand the impact of different types of variability on the performance of assembly and competing systems. Three sources of variability (processing time, unreliable workstations and imperfect yield) have been investigated here for the assembly system while the competing system was studied for processing time variability .For most of the cases we have also looked at ways of recovering capacity lost due to variability.

Our study reveals that with variability in processing time, throughput of an assembly system decreases with the c. v. and the percentage reduction in throughput is higher in PULL mode than in PUSH. Another interesting observation pertains to recovery of this lost capacity by the use of buffers. Percentage recovery appears to be inversely related to c. v. and is higher for PULL mode than for PUSH. In PUSH mode, the assembly station often acts as a natural buffer (with one unit being pushed up from a serial station while the other serial station is processing). Since, the marginal increase in throughput decreases with increasing buffer size hence an addition of, say, a single unit of buffer will yield a higher throughput in PULL (which has no natural buffer) than in PUSH mode. As observed by BPP, a large proportion of the lost capacity could be recovered with fairly small buffers. For small buffer sizes we observe that the WIP is slightly more than half the total buffer capacity in the system. Moreover. the higher is the c. v. the larger is the buffer size required to achieve a fixed percentage recovery level. We also explored the impact or increasing processing rate 00 percentage recovery.

In the case of unreliable stations we found that the throughput of the assembly system without buffers is dependent on *R*, the ratio of mean time to failure (U) and mean time to repair *(D).* For a given value of *R*, the throughput was the same for different values of *U* and *D*. However, with buffers, the throughput is de- pendent on *D*. We also observed that percentage recovery varies inversely with *D* for a given buffer size and processing rate, hence our conjecture that buffers in proportion to *D* will recover similar percentage of lost capacity .Our results confirmed this belief. For a given buffer size and *D* the percentage recovery changes marginally when *U* is varied.

One significant observation with respect to the impact of variability due to imperfect yield is that perfectly synchronized assembly systems are more affected by poor yield than those which show processing time variability as well. This should be reflected in the design of assembly systems and buffers. For 'serial stations"
in an assembly system, percent reduction in throughput due to less than perfect yield is found to *be* inversely related to the c. v. of processing time distribution.
We studied the pure competing system as well as configurations, which were combinations of competing, assembly and serial line systems. In all these cases a small butter size recovers a large fraction of capacity lost due to processing time variability .It is interesting to note that in configurations where more than one possible location exists for placing buffers, the choice of best location becomes critical for maximum recovery .In Configuration II, the center position for the buffer (after the -station which feeds the maximum number of stations) appears to be superior. Similarly, in Configurations III & IV buffers between two serial stations lead to higher recovery of throughput. Based on these observations.
We also provide insights to determine the best location buffers in production lines with competing system. Our simulation results provide an operations manager better insights into design of systems with and without buffer. The intuition obtained could be useful in future research on devising general models to explain the role of variability in production systems.

Acknowledgement

This research was supported by NSERC grant JGP0042150. The authors wish to thank the Department Editor Candace Yano, Richard Loulou and Dave? Pyke for several useful comments.

REFERENCES

[1] Altiok.T., and Perros,H.G.,"Open Networks of Queues with Blocking: Split and Merge Configurations," AIIE Transactions.18.251-261(1987).
[2] Altiok.T., and Stidham,S., " The Allocation of Interstage Buffer Capacities in Production Lines," IIE Transactions,15,292-299(1983).
[3] Alves, R., "Performance Evaluation of Series-Parallel Systems *Journal of manufacturing and Operations management*, 3, 224-250 (1990).

[4) Awate, P. G" and Sastry, B, L, N”.’Analysis and Decomposition of Transfer and Flow Lines," *Opsearch.* 24, 175-196 (1987
[5] Baker, K. R" Powell, S, G" and Pyke, D, F., "Buffered and Unbuffered Assembly Systems with Variable Processing Times." *Journal of manufacturing and Operations Management, 3,200-223 (1990),*
[6] Bhat, U. N" “Finite Capacity Assembly-Like Queues,” *Queueing Systems: Theory and Applications, I, 85 (1986),*
[7] Bhatnagar, R" and Chandra, P.,’Variability in Assembly and Competing Systems: Effect on Performance and Recovery Working Paper 91-02-18, Faculty of Management, McGill university, Montreal, Canada (1991),
8] Bonomi, F,. “An Approximate Analysis for a Class of Assembly Like Queues," *Queueing Systems: Theory and Applications,1, 89 (1987).*
[9] Buzacott, I. A,. "Prediction of the Efficiency of Production Systems without Internal Storage," IJPR, 6, 173-188 (1968).
[10] Buzacott, I, A" "Modelling Manufacturing Systems," *Robotics and Computer-integrated manufacturing, 2, 25-32 (1985),*
[11]Cornmault, C" and Dallery, y" "Production Rate of Transfer Lines without Intermediate Storage," *IIE Transactions, 22, I315-329 (1990).*
[12] Conway, R, W" Maxwell, W. L" McClain J, 0., and Thomas. L. J" .'The Role of Work-In-Process Inventory in Serial production Lines," *Operations Research, 36, 229-241 (1988).*
[13] Conway, R, W" Maxwell, W, L., McClain, J, 0" and Worona, S, L., User Guide to XCELL+, Scientific Press, Redwood City, CA (1986).
[15] Cumings, S., "Developing Integrated Tooling Systems: A case study of Garret Turbine Engine Company,” Proceedings of the fall Industrial Engineering Conference, Boston,MA(1986).
[16] Dallery, Y. and Gershwin.S.B.,” Manufacturing Flow Line Systems: A review of models and Analytical results,” Technical Report LMP-91-002. Department of Mechanical Eng., MIT, Cambridge, MA (1991).
[17] DiMascola, M., David,R., and Dallery, Y.,” Modeling and Analysis of Assembly Systems with Unreliable Machines and finite Buffers,”IIE Transaction,23,315-330(1991).

[18] Gershwin, S. B.,.’An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking," Operations Research, 35.291-305 (1987).
[19] Gershwin, S. B.,”Assembly/Disassembly Systems: An Efficient Decomposition Algorithm for Tree Structured Networks. ' , II£ Transactions, 23, 302-314 (1991),
[20] Harrison,J, M., " Assembly-Like Queues," Journal of Applied Probability, 10, 354 (1973),
[21] Hopp, W, J., and Simon, I. T., "Bounds and Heuristics For Assembly-Like Queues,.' Queueing Systems, 4, 137-156 (1989).
[22] Ignall. E" and Silver, A., .'The Output ofaTwo-StageSystem with Unreliable Machines and limited Storage," *AIIE* Transactions, 9, 183-188 (1977).
[23] Jafari, M. A" and Shanthikumar, J, G., .'Determination of Optimal Buffer Storage Capacities and Optimal Allocation in Multi- stage Automatic Transfer Lines, , , II£ Transactions , 21, 130-135 (1989),
[24] Kannarkar, U.S., "Manufacruring Lead Times, Order Release and Capacity Loading," Working Paper CMOM 89-04, Simon Graduate School of Business, University of Rochester (1989),
[25] Kubat, P" and Sumita, U., "Buffers and Backup Machines in Automatic Transfer Lines, International Journal of Production Research, 23, 1259-1270 (1985).
[26] Lee, H, S., and Pollock, S. M”. Approximation Analysis of Open Acyclic Exponential Queueing Networks with Blocking, 38, 1123-1134 (1990).
[27] Lipper, E. H., and Sengupta. B., Assembly-Like Queues with Finite Capacity: Bounds, Asymptotics and Approximations, Queueing Systems: Theory and Applications, I, 67 (1986),
[28] Muth, E, I., "Numerical Methods Applicable to a Production Line with Stochastic Servers," *Algorithmic methods in Probability,* M. F. Neuts (ed,), Studies in the Management Sciences. 7, 143-159 (1977),
[29] Perros, H. G., and Snyder, P. M., "A Computationally Efficient Approximation Algorithm for Analyzing Open Queueing Networks with Blocking," Manuscript, Computer Science Department, North Carolina State University (1986).

Rohit Bhatnagar is completing his doctoral studies in the Faculty of Management at McGi1I University and will be joining Nanyang Technological University in Singapore as a Lecturer, His research interests include multi-plant coordination and simulation and analysis of manufacturing systems.
Pankaj Chandra is an associate professor in the Faculty of Management at McGill University. He received a Ph.D. in Decision Sciences from the Wharton School of the University of Pennsylvania. His cur- rent research interests include manufacturing management, supply chain coordination, and technology management in the textile industry.