Reconfiguring a Remanufacturing Line at Visteon, Mexico

Sunder Kekre • Uday S. Rao • Jayashankar M. Swaminathan • Jun Zhang
Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
College of Business, University of Cincinnati, Cincinnati, Ohio 45221
The Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina 27599
A. B. Freeman School of Business, Tulane University, New Orleans, Louisiana 70118
sk0a@andrew.cmu.edu • uday.rao@uc.edu • msj@unc.edu • jzhang4@tulane.edu

This paper was refereed.

Visteon’s remanufacturing facility in Lamosa, Mexico was plagued with heavy fluctuations of supply and demand, leading to periods of severe capacity shortage. Management asked us to assess options for improving capacity. We developed a simulation-based line-configuration model that simultaneously considers line balancing and line length (number of production stations) to maximize the remanufacturing system’s effective throughput. We computationally analyzed the effect of processing-time variability on line-reconfiguration decisions, the effect of correlated task-processing times on throughput, and the marginal benefits of using dynamic line balancing. Based on the data we collected, we made recommendations for reconfiguring Visteon’s remanufacturing line. Management successfully implemented these changes, increased asset utilization, and reduced its planned new investments in capital equipment.

(Vacilities: equipment planning, capacity expansion. Manufacturing: performance, productivity.)

Visteon is a major supplier of integrated automotive systems, including power-train control, chassis, and electronics, to Ford Motor Company. Increasing competition and rapidly evolving technologies and market preferences are shortening product life cycles for such manufactured products as automobiles, computers, printers, copiers, and cameras. Simultaneously, improved materials, product designs, and manufacturing processes are increasing the reliability and life spans of some of the components of these products. Growing environmental concerns about the magnitude of waste products generated by the manufacturing sector along with cost considerations have prompted some firms, such as Visteon, to collect reusable product modules and refurbish them for sale in the aftermarket. Environmental management is one of Visteon’s highest corporate priorities (www.visteon.com).

Visteon uses the term remanufacturing (or reman) for transforming used units (cores) into refurbished units that satisfy exactly the same quality standards as new units. Visteon’s reman process is labor intensive and highly variable, with processing times depending on the age and amount of wear and tear in the input core. Visteon’s plant in Lamosa, Mexico remanufactures product families, such as rack-and-pinion steering gears, recirculating ball-nut gears, and power-steering pumps for several models of Ford cars, trucks, and recreational vehicles. We focus on the rack-and-pinion (R and P) gear product family, although we conducted similar analyses of other reman products made at Lamosa. The mixed-model R and P reman line is structured like a repair facility: Cores are first stripped of worn-out (nonreusable) parts, then cleaned, inspected visually, tested for straightness and cracks, machined, assembled, tested for functionality,
Figure 1: The rack-and-pinion production line consists of a series of operations performed at various workstations. (The stations in bold are bottlenecks. We focused on the four-station assembly line.)

and finally, packed and shipped. The R and P line has few product variants with the top eight product families accounting for over 80 percent of the demand. The quality of input cores varies; this variability often results in loss of throughput.

Visteon’s Dilemma

Visteon’s managers hesitated to make additional capital investments to overcome the adverse impact of variability and struggled to identify the best option for reconfiguring the plant. A group of process engineers thought they could achieve gains in throughput by reconfiguring the reman line, but they could not quantify the gains. Another group favored duplicating the existing line to increase capacity. The managers asked us to resolve this dilemma. After visiting the production facility and collating results from time studies and available data, we identified cleaning, washing, assembly, and functional testing workstations as the bottlenecks limiting throughput. We learned that management had an elaborate plan to improve three of the bottleneck stations; this plan had been approved and was to be implemented soon. Consequently, the last assembly station (Assembly D) was the only bottleneck remaining to be addressed. We studied Visteon’s four-station assembly system because we could alleviate the Assembly D bottleneck by changing the assignment of tasks and workers to the different assembly stations.

Visteon operated its four-station assembly system in an asynchronous mode and had no storage buffers between the stations because of the firm’s zero-inventory policy. Further, workers were dedicated to particular assembly stations and were trained to operate fixtures and machines corresponding to their assigned tasks. Workers tested the core for various defects, and this inspection determined the processing-time requirements at the four assembly stations. The processing times at a station varied between 500 and 750 seconds, and task times at these stations could be positively or negatively correlated.

Our first step was to identify the key drivers that affected throughput in the reman line. Production capacity in general at a reman line depends on several factors, such as line and process design (layout, number of workstations operating in series or parallel, use of interstation buffers and other equipment), line operation (balancing, worker training), and characteristics of the operating environment (processing time variability and correlation, core age and wear). Clearly, Visteon could not reduce the variability of processing times because it had limited control over the quality of cores it received. Given these constraints, Visteon management wanted answers to the following questions from our research team.

1. Should Visteon use a number of short lines or a few long lines? What difference in throughput would result?

2. What improvements could Visteon make by dynamically assigning tasks to workstations?

The dynamics of a remanufacturing line are quite different from those of a traditional manufacturing line at Visteon. First, Visteon often controls its traditional lines using material requirements planning (MRP). That is, core input is driven by a production schedule for end products. Planners convert the schedule to demands for input components using bill-of-materials data on usage quantities, which specify the number of units of each component required to make one unit of each end product. In our remanufacturing (repair) environment, usage quantities of components are uncertain because component consumption depends on the quality of the cores. Good quality cores need fewer components replaced than old, worn-out cores. This uncertainty often precipitates a mismatch between supply and demand. Second, planners know the processing times in traditional lines fairly accurately. In contrast, reman lines exhibit varying processing times in repair and
finishing. Further, these times are not known until the core is tested.

To address these issues, we develop an optimization and simulation model that we deploy to recommend improvements to Visteon’s reman line.

Literature Review

In its remanufacturing, Visteon faces new operations management challenges in product and process design for remanufacturing, reverse logistics for core retrieval, and tactical plant loading to match demand with production under uncertainty. In this paper, we focus on production planning and control (PPC) and line design. Several factors complicate the use of traditional PPC methods in remanufacturing, including stochastic routing, probabilistic use of materials and components, and variable processing times. Guide (2000) classifies remanufacturing research into reverse logistics, forecasting, PPC, and inventory. Ferrer and Whybark (2001), Haynsworth and Lyons (1987), McConocha and Speh (1991), and Panisset (1988) focus on reman and inventory; Panisset (1988), Perry (1991), and Sprow (1992) consider scheduling for reman.

Researchers have used several quantitative approaches. Fleischmann et al. (1997) reviewed models for reverse logistics; Inderfuth et al. (2001) modeled stochastic inventory management for reverse logistics; van der Laan et al. (1999) analyzed inventory control policies for hybrid systems with both reman for the aftermarket and traditional original equipment manufacturing; Toktay et al. (2000) used a closed-queuing framework for inventory replenishment; Ketzenberg et al. (2002) used a queuing model and simulation to study a mixed assembly-disassembly line for remanufacturing; Ferrer (2003) explored the interactive effects of supply-delivery lead times and early detection of core yield; and Ferrer and Swaminathan (2002) studied pricing and production decisions for new and reman products.

Researchers have extensively studied the design, analysis, and control of a serial production line system with variable operation times (Buzacott and Shanthikumar 1993; Hopp and Spearman 2000, Chapters 7–9). The issues considered include (1) balancing the line by allocating different tasks to workstations (Hillier and Boling 1966, Kao 1976, Carraway 1989, Zavadlav et al. 1996, Bartholdi et al. 1999); (2) determining the number and position of buffer storage spaces between workstations in the production line (Conway et al. 1988, So 1997, Yamashita and Altiok 1998); (3) choosing the geometric layout, for example, straight line, U line, Y line (Chand and Zheng 2001); (4) selecting the line length, that is, the number of workstations to use in a production line (Hillier and Boling 1966, Chase 1975).

Models for Line Reconfiguration

We modeled Visteon’s reman assembly operation as a serial production line in which remanufacturing the product from the core requires $M$ tasks. The processing time of task $m$ is a random variable denoted by $\xi_m$, where $\xi_1, \ldots, \xi_M$ may be correlated. The tasks are ordered so that, for $m = 1, \ldots, M - 1$, task $m$ must be completed before task $m+1$ can start. The serial line is segmented into $N$ stations, with each station assigned a subset of the $M$ tasks, to be performed by a single team of workers. Visteon’s current assembly process for the R and P line has four such stations. Before a core enters the assembly line, workers test the core and develop a list of tasks required in assembly, along with a fairly accurate estimate of task-processing times. Consequently, we can assume that all processing times for tested cores are known with certainty. For Visteon’s reman operation, we had to decide how many production stations to have in each assembly line and the number of copies of this line. To effectively compare different combinations of number of stations per line, $N$, and number of lines, $L$, we keep $L \times N$ constant. That is, we fix the total number of worker teams (stations) to $LN$. For instance, with six workstations permitted, we examine whether it is better to have three two-station production lines or two three-station production lines or six short one-station lines or one long six-station line.

Alternative line configurations affect tooling, worker-training, and worker-retention costs. At Visteon, tooling costs were of little importance because we were focusing on a manual assembly process.
located in Mexico. Further, the large backlog of profitable customers suggested that a modest investment in the equipment needed for reconfiguration could have rapid payback. The interplay of human resource factors, such as labor pay, turnover, and training times, was important; however, we considered the hypothesized technical advantages of shorter lines—lower balance losses and higher risk pooling—most important. Visteon was willing to address the human resource issues only if we could demonstrate that shorter lines would produce significant throughput benefits. Further, management required a threshold improvement in throughput of more than 20 percent to consider any proposed changes to the system.

Visteon’s reman operations had interesting characteristics in terms of modeling and analysis. First, to maximize throughput, we needed to improve both line balancing and line configuration. Second, the randomness in processing durations stemmed from uncertainty about the conditions of retrieved cores. However, once workers tested a specific core for wear, they could determine its processing durations accurately. Third, Visteon’s operations policy set limits, such as zero buffer inventory. To quantify the throughput benefits of alternative line configurations, we had to model two related aspects. First is the choice of the configuration \((L, N)\), that is, the number of parallel lines and the number of stations in each line, that maximizes the production throughput per worker team. We call this the line-configuration problem (LCP). Once we determine the line configuration, we have to coordinate each line by effectively assigning the \(M\) processing tasks to the \(N\) workstations so as to balance workloads. We call this the line-balancing problem (LBP). Clearly, LCP and LBP are interrelated. We developed a method of optimizing the two problems simultaneously. We present details of the formulation for LCP and LBP in the Appendix.

In practice, for a given number of stations and prespecified task allocations, the throughput of the production line still depends on how the firm controls the remanufacturing process, the amount of buffer storage space between stations, and how the firm releases cores into the remanufacturing line. For our analysis, we assumed Visteon had an unlimited supply of recycled cores and unlimited demand for remanufactured products. This is a standard assumption in modeling production lines when analysts have insufficient information on future supply and demand. Further, they can model systems with supply and demand uncertainty as lines with unlimited supply and unlimited demand by adding a dummy workstation at the beginning of the line to model supply and another station at the end of the line to model demand. We did not add such dummy stations because we did not want the evaluation of line reconfigurations to be affected by uncertain factors extraneous to the line itself. Similarly, we did not insert buffers between stations because many automotive remanufacturing lines, including Visteon’s, use lean manufacturing with zero buffers. However, we did investigate the impact of the zero-buffer restriction. Past research has indicated that a few strategic buffers can significantly improve line performance, and we wanted to demonstrate the potential benefits to management.

There are two ways of controlling material flow: synchronous or asynchronous. A serial line is synchronous if each task is held at its workstation until all tasks in the line are ready to proceed to their next workstation. A serial line is asynchronous if a part completed at one station enters the next station as soon as this successor workstation is available (Chand and Zheng 2001). Visteon’s reman line operated in an asynchronous fashion. It is not easy to evaluate the performance of asynchronous serial systems with random processing times and finite buffers. Currently no known closed-form expressions exist for the expected throughput of Visteon’s four-station assembly line. Further, standard results, such as those in Buzacott and Shanthikumar (1993) do not directly apply to our systems. As a result, we could not answer all the questions the managers posed with the knowledge available. We first explored the problem theoretically.

Our analysis suggested that our first task was to understand the distribution of processing times in the \(R\) and \(P\) line. We visited the facility and conducted time studies over a number of days. We also obtained data from the firm for the past several months on the mean and variance of the processing times. After a statistical test, we found that the data did not match an exponential distribution. However, it fitted well
Reconfiguring a Remanufacturing Line at Visteon, Mexico

with a gamma distribution. To provide insights on the questions managers had raised and to quantify the benefits of reconfiguring the line under different operating conditions, we needed to perform an extensive computational study.

Computational Study of the Visteon Line

We performed a computational study based on data obtained from our experience at Visteon’s automotive parts remanufacturing facility. (We disguised the numbers to maintain confidentiality.) In our base experiments, we modeled a system in which each job requires $M = 4$ tasks with each task-processing time assumed to be independent and identically distributed (gamma or normal). We chose $M = 4$ and identical distributions for processing times to reflect Visteon’s actual task environment. Our computational results with nonidentical processing times yielded similar qualitative insights. We chose a mean processing time of 10 minutes and varied the coefficient of variation, $cv$ (standard deviation/mean), for gamma distributed processing times from 0.1 to 1.0 in steps of 0.1. For normal distributed processing time, we set $cv$ to 0.3 for each task and we varied the correlation ($\rho$) between the processing times of different tasks from $-0.3$ to $+0.3$.

To conduct the computational experiments, we used a discrete-event simulator coded in C programming language. For each configuration of the line, we generated 12,000 jobs; we used the first 2,000 jobs to initialize the system. We calculated throughput based on the time interval between the entering time of job #2,001 and the departure time of job #12,000. For these parameters, the standard error in the simulation estimate of throughput is of the order of a fraction of a percent for normally distributed processing times and of the order of a few percent for gamma distributed processing times. We conducted the simulation experiments on a SUN Sparc10 workstation.

Managerial Insights

The Effects of Line Configuration

We compared the throughput of one-, two-, three-, and four-station lines. Because each station has one worker team, longer lines effectively have more workers. To generate comparable results, we had to keep the workforce level constant. Because the least common multiple of 1, 2, 3, and 4 is 12, we fixed the number of worker teams to 12. Thus, we compared the performance of 12 one-station lines, six two-station lines, four three-station lines, and three four-station lines (this normalizes the results with respect to the number of worker teams). We determined the effects of line configuration under different values of the coefficient of variation for the task-processing time (Figure 2). The throughput behavior for lines with $N = 1, 2, 4$ is different from that of a line with $N = 3$. This may be explained by the the fact that the latter case ($N = 3$) is the only one in which the line is not balanced.

For the test data ($N = 1, 2, 4$), shorter balanced lines always yield higher system throughput than longer lines (Figure 2). When per-station processing-time variability increases, the throughput of the shorter lines decreases more slowly than the throughput of longer balanced lines. Thus, balanced shorter lines are more robust with respect to variability.

One explanation for these observations is that serial lines without buffers lose throughput because of workstation starvation and blocking, which increases with line length and with processing-time variability. Conway et al. (1988) had similar results, except that they considered lines with identical workstations (with uniform $0, 1$ processing times) and no line balancing. In our model, the workstations need not be identical. Because we consider a single job type with $M$ tasks, workstations corresponding to shorter lines will process more tasks. This results in risk pooling (reduced effective coefficient of variation of processing time), which reduces the chance of blocking and starvation.

As $cv$ increases, the throughput of the longer balanced four-station line decreases faster than the throughput of the shorter unbalanced three-station line. However, the longer line has higher throughput at low $cv$ (Figure 2). Consequently, there is a threshold $cv$ below which the four-station line’s throughput dominates the three-station line’s throughput. Beyond this threshold, the unbalanced three-station line has higher throughput.
Hillier and Boling (1966) showed that an $N$-station unbalanced line can have higher throughput than an $N$-station balanced line. Our experiments revealed that an unbalanced shorter line (with three stations) may yield higher throughput than a balanced longer line when its processing-time variability is higher. This finding suggests that the deterioration in throughput due to line unbalance and processing-time variability may sometimes be overcome by reducing the line length.

**The Effect of Line Balancing**

To quantify the improvement in expected throughput resulting from use of dynamic line balancing, we compare the throughput of six two-station lines versus four three-station lines, under both static and dynamic line balancing (Figure 3). Because each product consists of four tasks, there is no difference between dynamic and static line balancing for the one-station line (in which all tasks are assigned to the single-station under both line-balancing rules) and the four-station line (in which each task is assigned to a different station). Given only $M = 4$ tasks, we simply solve the line-balancing problem (LBP) for $N = 2$ and $N = 3$ by enumeration. From these experiments, we have three observations:

1. For the same $cv$ and $N$, dynamic line balancing (DLB) always yields higher throughput than static line balancing (SLB); the improvement is greater for the three-station unbalanced line than for the two-station balanced line. DLB’s effectively balancing the line causes this improvement.

2. For the balanced two-station line, the improvement in throughput under DLB is greater when the processing-time variability is higher; whereas for the unbalanced three-station line the throughput improvement is greater when the processing-time variability is lower with $cv$ values near 0.2. Opportunities to improve throughput by line balancing stem from two sources: (1) line unbalance...
Figure 3: This figure shows that, for each coefficient of variation \((cv)\) of task processing time, use of dynamic line balancing (DLB) instead of static line balancing (SLB) improves system throughput. This improvement is higher for the unbalanced three-station line than for the balanced two-station line.

(3) For correlation \(\rho\) between \(-0.3\) and \(+0.3\), the throughput improvement under DLB decreases in a near-linear fashion as the processing times become more positively correlated, largely because, under positive correlation, there is less opportunity for effective, reactive (dynamic) line balancing.

DLB may be more useful for balanced lines (same expected processing time per station) with high variability or unbalanced lines with low to mid variability (Figure 3).

As part of our sensitivity analysis, we conducted additional experiments to study the benefits Visteon could expect by permitting inventory buffers between assembly stations, and the throughput behavior of lines with more than four tasks (in particular, we considered \(M = 12\) and \(M = 24\), with number of stations, \(N = 1, 2, 3, 4, 6, 12,\) and \(24\)). By buffer space, we mean a storage capacity of one unit. By total buffer size, we mean the total number of buffer spaces in the entire line. Our experiments suggest that (1) to achieve the same percentage improvement in throughput, fewer buffer spaces are required for higher processing time \(cv\); (2) the improvement in throughput with total buffer size does not always exhibit diminishing returns in that the improvement is sensitive to the optimal deployment of buffer spaces.
within the line; (3) the throughput of a four-station line with a total of five to six buffer spaces is similar to that of a two-station line with zero buffers; (4) the throughput of balanced lines decreases as the number of stations increases, even for systems with more than four tasks \((M > 4)\); and (5) given the same number of stations, the system with the most tasks sees the greatest benefits (in terms of percent improvement in throughput) from changing the single long line to multiple shorter lines with the same total number of stations.

**Throughput Improvement at Visteon: The Big Surprise**

One of the main objectives of our research was to provide Visteon with insights on the qualitative and quantitative benefits of line reconfiguration. Visteon management was interested in line reconfigurations that would improve throughput by more than 20 percent (Table 1).

The managers were pleasantly surprised that they could improve throughput by as much as 90 percent without significant capital investments (when \(cv \geq 0.5\)). However, the data we obtained from Visteon indicated that the R and P reman line at Lamosa had processing-time \(cv\) between 0.3 and 0.4. Even in this range, our study indicated that the firm could improve throughput by as much as 35 percent by converting to a single-station assembly line.

<table>
<thead>
<tr>
<th>New Configurations</th>
<th>Processing-time (cv = 0.3–0.4)</th>
<th>Processing-time (cv = 0.5–1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-station line</td>
<td>25–35</td>
<td>44–90</td>
</tr>
<tr>
<td>Two-station line with DLB</td>
<td>—</td>
<td>25–45</td>
</tr>
<tr>
<td>Two-station line with SLB</td>
<td>—</td>
<td>23–40</td>
</tr>
</tbody>
</table>

*Table 1: For different processing-time \(cv\) values, we show the percent improvement in throughput for line configurations that yielded an improvement of more than 20 percent (relative to the current four-station system). We excluded the case with \(cv < 0.3\) because it did not yield the required improvement.*

Our Recommendations and Challenges at Lamosa

Our analysis helped Visteon to identify the best design for its product lines. The assembly stations in the R and P line with a task-processing time \(cv\) of 0.3 are ideally suited for transformation to single-station lines (to improve throughput by at least 25 percent). Further, a single-station line does not need balancing. Its only drawback is that workers would need cross training, but plant managers favored cross training because it would mitigate the detrimental effects of worker absenteeism and make one worker team directly responsible for a specific product’s assembly quality. In addition, by using single-station assembly and cross training, managers would gain the flexibility to redirect assembly workers to accommodate the shop floor workload. Our study helped the firm in three ways. First, it provided managers with qualitative insights on line reconfiguration. Second, it enabled Visteon to quantitatively assess the benefits of alternative configurations. Finally, it gave managers target performance levels (in terms of throughput) for benchmarking their reconfigured lines. Based on our recommendations, the plant managers have reconfigured the four-station assembly line into four single-station lines. Further, they are considering a similar reconfiguration of the other reman lines at Visteon Lamosa.

During our study of the reman line, we realized that Visteon had several other issues that needed to be studied. In particular, Visteon’s severe shortage of capacity created an incentive for customers to place fictitious orders to ensure that they got part of the capacity. On closer analysis of the data, we realized that the capacity situation was not as bad as it seemed. Currently managers at Lamosa are evaluating strategies to reduce the apparent variability in demand by giving customers incentives to inform them of their true demand instead of placing fictitious orders. Second, Visteon’s lack of a good production-planning tool for the remanufacturing line made it very difficult for it to coordinate demand and supply. Based on our recommendation, the firm is modifying planning tools for remanufacturing environments to leverage the reconfigurations we recommended. Our study strengthened
Visteon’s commitment to environmentally conscious manufacturing.

Appendix

The Line Configuration Problem (LCP)

Let $T(N)$ denote the cycle length of an $N$-station line, which corresponds to the maximum total work assigned to any station. Then, the throughput of the $N$-station line may be measured as $TH(N) = 1/T(N)$. Because we are interested in the throughput of a system with $L$ identical lines and $N$ stations per line, and we vary $N$ while keeping $LN$ fixed, we focus on the throughput per station, which is $TH(N)/N$. This quantity when multiplied by $LN$ (which remains constant) yields system throughput.

The line-configuration problem of maximizing the expected throughput per station may be stated as follows:

\[
\text{LCP: } \max_{N_b \leq N \leq M} \frac{1}{N} E[TH(N)],
\]

where $N_b$ denotes a prespecified minimum number of stations below which the line will become impractical to run. Because the objective does not, in general, possess good properties, such as concavity or monotonicity, this problem is complex to solve.

The Line Balancing Problem (LBP)

Among the many different ways of balancing a line, we focus on two: static line balancing (SLB) and dynamic line balancing (DLB). Under static line balancing, one allocates tasks to different stations before testing each core, with line balancing based on expected task-processing times ($E\xi_1, \ldots, E\xi_M$). Under dynamic line balancing, one allocates tasks to stations for each individual core, after testing cores and determining the processing time of all tasks. In this case, the line balancing is based on each realization of $(\xi_1, \ldots, \xi_M)$. Visteon could implement such a scheme if, after initial testing of each core, it attached a process-routing slip indicating the tasks to be executed and assigning required tasks to the different stations of the assembly system.

Let $x_{mn}$ be 1 if task $m$ is assigned to station $n$ and zero otherwise. Let $\eta_n$ denote the total workload assigned to station $n$, which is equal to the sum of the processing times of all tasks assigned to $n$. The intermediate variable $T$ will correspond to the highest station workload, $\max_n(\eta_n)$. With this notation, we can formulate the line-balancing problem of allocating the $M$ tasks to $N$ stations to minimize the cycle length as follows:

\[
\text{LBP: } T(N) = \min_{x_{mn} \in \{0, 1\}} T \quad (1)
\]

s.t.

\[
\sum_{n=1}^{N} x_{mn} = 1 \quad \text{for all } m, \quad (2)
\]

\[
\sum_{m=1}^{M} x_{mn} \geq \sum_{i=1}^{i} x_{m+i} \quad \text{for all } i = 1, \ldots, N, \quad m = 1, \ldots, M-1, \quad (3)
\]

\[
\eta_n = \sum_{m=1}^{M} x_{mn} \tilde{\xi}_m \quad \text{for all } n, \quad (4)
\]

\[
T \geq \eta_n \quad \text{for all } n, \quad (5)
\]

where $\tilde{\xi}_m = E[\xi_m]$ for static line balancing and $\tilde{\xi}_m = \xi_m(\omega)$ for dynamic line balancing, with $\xi_m(\omega)$ denoting a realization of processing time. Constraints (2) ensure that every task is assigned to some station, and constraints (3) ensure that task $m$ is completed at an earlier station (with lower index $n$) and task $m+1$ is assigned to a later station in the flow line. Constraints (4) define $\eta_n$ to be equal to the total amount of work assigned to station $n$. Constraints (5), along with the objective, ensure that $T = \max_n(\eta_n)$. This $T$ corresponds to the minimum feasible cycle length. In the remainder of the paper, whenever we refer to $TH_{\text{asy}}(N, \eta)$, we will implicitly assume that $\eta$ corresponds to tasks being allocated to the $N$ stations so as to balance the line, either using SLB or using DLB, depending on the context. Our main analytical results are as follows:

Result 1. For an asynchronous production line with four tasks having exponential and identically distributed task times (with mean 1/$\mu$) under static line balancing, the expected throughput per station is monotonically decreasing in the number of stations, that is,

\[
TH_{\text{asy}}(1, \eta) / 1 \geq TH_{\text{asy}}(2, \eta) / 2 \geq TH_{\text{asy}}(3, \eta) / 3 \geq TH_{\text{asy}}(4, \eta) / 4.
\]
Proof. When we use only one station, we assign all tasks to this station. In this case, the processing time of the station is a four-Erlang with mean $4/\mu$. Therefore, $TH_{\text{asy}}(1, \eta) = \mu/4$. Comparing the throughput per station under different numbers of stations using Lemmas 1, 3, and 4 (proved below) yields the required result. □

Lemma 1. For a four-station asynchronous serial line with independent exponential processing time (mean $1/\mu$) on each station, the expected throughput is $TH_{\text{asy}}(4, \eta) = 5\mu/14$.

Proof. Let $N_i(t)$ denote the number of jobs at station $i$ for $i = 2, 3, 4$ at time $t$. Then $\{N_2(t), N_3(t), N_4(t), t \geq 0\}$ is a Markov process on the state space $S = \{(n_2, n_3, n_4): n_i = 0, 1\}$. Let $p(n_2, n_3, n_4)$ be the steady state probability that the system is in state $(n_2, n_3, n_4)$. The balance equations for this four-station line are given by the following:

$\mu p(0, 0, 0) = \mu p(0, 0, 1),$

$\mu p(0, 0, 0) + \mu p(0, 1, 0) = \mu p(1, 0, 0),$

$\mu p(1, 0, 0) + \mu p(1, 1, 0) = (\mu + \mu)p(0, 1, 0),$

$\mu p(0, 1, 0) = (\mu + \mu)p(0, 1, 1),$

$\mu p(0, 1, 1) + \mu p(1, 1, 0) = (\mu + \mu)p(1, 0, 1),$

$\mu p(1, 0, 1) = (\mu + \mu)p(1, 1, 0),$

$p(0, 0, 0) + \cdots + p(1, 1, 1) = 1.$

Solving the above linear system, we have

$TH_{\text{asy}}(4, \eta) = \mu \times \text{Prob}[\text{Station 4 is busy}]$

$= \mu \times \sum_{i,j} p(i,j,1)$

$= 5\mu/14.$ □

Lemma 2 (Muth and Alkaff 1987). Consider a three-station asynchronous serial line with $\eta_i$, $i = 1, 2, 3$, denoting the total processing time at each station. Let the distribution functions of $\eta_1$ and $\eta_3$ have the form

$F_{\eta_1}(t) = 1 - \sum_{i=1}^{m} \beta_i \exp(-\alpha_i t)$ for $t \geq 0,$

$F_{\eta_3}(t) = 1 - \sum_{j=1}^{n} \delta_j \exp(-\gamma_j t)$ for $t \geq 0,$

where $\alpha_i$ and $\gamma_j$ are sets of distinct positive real numbers, $\beta_i$ and $\delta_j$ are real numbers that satisfy $\sum_{i=1}^{m} \beta_i = 1$ and $\sum_{j=1}^{n} \delta_j = 1$. If the distribution of $\eta_2$ has a Laplace transform, $F_{\eta_2}(s)$, then the expected throughput ($TH$) of the line is given by

$$\frac{1}{TH} = E[\eta_2] + \sum_{i=1}^{m} \beta_i F_{\eta_2}^{*}(\alpha_i)$$

$$+ \sum_{j=1}^{n} B_j \left[ F_{\eta_2}^{*}(\gamma_j) - \sum_{i=1}^{m} \beta_i F_{\eta_2}^{*}(\alpha_i + \gamma_j) \right],$$

where $B_j, j = 1, \ldots, n$ are obtained by solving the following system:

$$B_j = \delta_j - \sum_{i=1}^{m} \beta_i \gamma_j A_i/(\alpha_i + \gamma_j)$$ for $j = 1, \ldots, n,$

$$A_i = \beta_i \alpha_i F_{\eta_2}^{*}(\alpha_i) - \sum_{j=1}^{n} \beta_i \alpha_i B_j F_{\eta_2}^{*}(\alpha_i + \gamma_j)$$ for $i = 1, \ldots, m.$

Lemma 3. For an asynchronous production line with $M = 4$ tasks, if task $m$’s processing time $\xi_m$ is an independent exponential random variable with mean $1/\mu$, the throughput of a three-station line is $TH_{\text{asy}}(3, \eta) = 40\mu/97$.

Proof. When we use three stations, we have three possible ways to allocate four tasks to the three workstations, depending on which station is assigned two tasks. Let LB$i$ denote the case in which station $i$ is assigned two tasks. Due to the reversibility of a serial line, LB1 will yield the same throughput as LB3 (Muth 1973). Therefore, we need to evaluate only the throughput of LB2 and LB3.

For LB2, the processing times for stations 1 and 3 are exponential random variables with mean $1/\mu$; hence $\alpha_1 = \gamma_1 = \mu$ and $\beta_1 = \delta_1 = 1$. The processing time for station 2 is two-Erlang with mean $2/\mu$; therefore $F_{\eta_2}^{*}(s) = \mu^2/s(\mu + s)^2$. We can evaluate the throughput of this line using the formula in Lemma 2, which yields

$$TH_{\text{asy-LB2}}(3, \eta) = 40\mu/97. \quad (6)$$

For LB3, the processing times for stations 1 and 2 are exponential random variables and the processing time for station 3 is a two-station Erlang distributed random variable with mean of $2/\mu$. For this
line, we cannot apply the general formula provided by Lemma 2 because the processing time of station 3 does not belong to the special phase-type distribution. However, we can evaluate the throughput by following the procedure used to derive the formula in Lemma 2.

Following the notation of Muth and Alkaff (1987), let $H$ denote the holding period of a job at station 1 where the holding period is the sum of the processing time plus the blocking period. Let $I$ denote the idle period at station 2 following a departure. Let $R_3$ denote the residual service period of a job at station 3 in Lemma 2. Let $f_\eta(t)$ and $F_\eta(x)$ denote, respectively, the PDF and CDF of a random variable $\xi$. Note that $f_\eta(t) = \mu e^{-\mu t}$ and $F_\eta(x) = \mu x e^{-\mu t}$. Then,

\begin{align}
F_s(t) &= 1 - \int_{x=0}^{\infty} F_\eta(x) R_3(x) \mu e^{-\mu(x+t)} dx, \\
F_R(t) &= 1 - \int_{x=0}^{\infty} F_\eta(x) \mu x e^{-\mu(x+t)} dx.
\end{align}

To simplify the above equations, let $A = \int_{x=0}^{\infty} F_\eta(x) R_3(x) \mu e^{-\mu x} dx$, $B = \int_{x=0}^{\infty} F_\eta(x) \mu^2 x e^{-\mu x} dx$, and $C = \int_{x=0}^{\infty} F_\eta(x) \mu^2 e^{-\mu x} dx$. Then, Equations (7) and (8) may be written as

\begin{align}
F_s(t) &= 1 - A e^{-\mu t}, \\
F_R(t) &= 1 - B e^{-\mu t} - C t e^{-\mu t}.
\end{align}

Plugging (10) into the definition of $A$ and using the fact that $\eta$ is an exponential random variable, we obtain

\begin{equation}
A = \int_{x=0}^{\infty} (1 - e^{-\mu x})(1 - B e^{-\mu t} - C t e^{-\mu t}) \mu e^{-\mu x} dx.
\end{equation}

Similarly, plugging (9) into the definitions of $B$ and $C$ yields

\begin{align}
B &= \int_{x=0}^{\infty} (1 - A e^{-\mu x}) \mu^2 x e^{-\mu x} dx, \\
C &= \int_{x=0}^{\infty} (1 - A e^{-\mu x}) \mu^2 e^{-\mu x} dx.
\end{align}

Solving the system of (11), (12), and (13), we get $A = 7/32$, $B = 121/128$, $C = 57 \mu / 64$. Therefore, by (10),

\begin{equation}
F_R(x) = 1 - \frac{121}{128} e^{-\mu x} - \frac{57}{64} \mu t e^{-\mu t}
\end{equation}

and hence

\begin{equation}
F_H(x) = \frac{F_R(x)}{F_\eta(x)} F_\eta(x) = (1 - e^{-\mu x})(1 - \frac{57}{64} \mu x e^{-\mu x}).
\end{equation}

It follows that the throughput under LB3 is

\begin{equation}
TH_{\text{asy-m}}(3, \eta) = \frac{1}{E[H]} = \frac{48 \mu}{121}.
\end{equation}

Comparing Equations (6) and (14), we see that $TH_{\text{asy-m}}(3, \eta) = 40 \mu / 97$. □

**Lemma 4.** For an asynchronous production line with $M = 4$ tasks, if task $m$'s processing time $\xi_m$ is an independent exponential random variable with mean $1/\mu$, the throughput of a two-station line is $TH_{\text{asy-m}}(2, \eta) = 4 \mu / 11$.

**Proof.** When we use two stations, one way of line balancing is to allocate two tasks to each station. Then, the processing time for each station is two-Erlang with mean $2/\mu$. Because the throughput of a two-station asynchronous line is the same as that of a two-station synchronous line, the throughput under this line balancing is $1/E[\max[\eta, I]] = 4 \mu / 11$. The other way of balancing the line is to allocate three tasks to one station and the fourth task to the other station. Because the expected processing time for the station with three tasks is $3/\mu$, the throughput of this line is less than $\mu / 3 \leq 4 \mu / 11$. Therefore, $TH_{\text{asy-m}}(2, \eta) = 4 \mu / 11$. □

Because the processing times are identical in Result 1, the four-station line is balanced (with the same expected processing time at each station). However, the three-station line is not balanced, even though its throughput per station is higher. The managers saw that an unbalanced line with fewer stations could have higher throughput than a perfectly balanced line with more stations. This observation is related to that of Hillier and Boling (1966), who observed empirically that unbalancing a line while keeping the number of stations constant can sometimes improve throughput in the presence of processing-time variability. While Result 1 holds for the above four-task system with i.i.d. exponential task durations, it may not hold in general.
Result 2. For an asynchronous production line with three tasks having independent and identically distributed task times (with mean $1/\mu$) that are either (1) exponential, or (2) uniform, the expected throughput per station under static line balancing is not monotonically decreasing in the number of stations, that is,

$$TH_{\text{asy}}(2, \eta)/2 \leq TH_{\text{asy}}(3, \eta)/3 \leq TH_{\text{asy}}(1, \eta).$$

Proof. (1) When we use only one station, we assign all tasks to this station. In this case, the processing time of the station is a three-Erlang with mean $3/\mu$. Therefore, $TH_{\text{asy}}(1, \eta) = \mu/3$. To evaluate $TH_{\text{asy}}(2, \eta)$, we need to consider only line-balancing cases in which two tasks are allocated to the first station and the third task is allocated to the second station (due to reversibility). Then, the processing time for station 1 is a two-Erlang with mean $2/\mu$, and the processing time for station 2 is exponential with mean $1/\mu$. Because the throughput of a two-station asynchronous line is the same as that of a two-station synchronous line, the throughput under this line balancing is $1/E[\max(\eta_1, \eta_2)] = 4\mu/9$, which may be verified through lengthy but straightforward algebra. Therefore, $TH_{\text{asy}}(2, \eta)/2 = 2\mu/9$.

We evaluate $TH_{\text{asy}}(3, \eta)$ using the formula in Lemma 2 with $\alpha_1 = \gamma_1 = \mu, \beta_1 = \delta_1 = 1$, and $F_{\eta_i}(s) = \mu/(\mu + s)$. It is easy to verify that $TH_{\text{asy}}(3, \eta)/3 = 3\mu/11$. Comparing $TH_{\text{asy}}(3, \eta)/3 = TH_{\text{asy}}(2, \eta)/2$, and $TH_{\text{asy}}(1, \eta)$, we complete the proof of (1).

Result 2(2) is proven similarly. When we allocate tasks 1 and 2 to the first station, the processing time of station one is a triangle distributed random variable between zero and two with mean one; and the processing time of station 2 is a uniform random variable between zero and one. In this case, $TH_{\text{asy}}(2, \eta) = 1/E[\max(\eta_1, \eta_2)] = 0.8$. For a single-station line, $\eta_i$, the processing time of the station is the sum of three independent uniform random variables between zero and one. In this case, $TH_{\text{asy}}(1, \eta) = 1/E[\eta_1] = 2/3$. By Lemma 5 (proved below), $TH_{\text{asy}}(3, \eta) = 1.341$. We obtain the required result by comparing $TH_{\text{asy}}(N, \eta)/N$ for $N = 1, 2, 3$. □

Lemma 5. Consider an asynchronous production line with $M = 3$ tasks. If task $m$’s processing time $\xi_m$ is an independent uniform random variable, then the expected throughput of a three-station asynchronous line is 1.341.

Proof. Similar to Lemma 3, from Equations (7) and (8), we have

$$F_3(t) = 1 - A, \quad F_{R_3}(t) = 1 - B,$$

$$A = \int_{x=0}^{1-t} F_{\eta_2}(x)F_{\eta_3}(x) \, dx,$$

$$B = (1 - A)/(1 - t).$$

Solving the above system, we obtain,

$$A = -\frac{-t + 2t^2 - t^3}{1 + 3t - 3t^2 + t^3} \quad \text{and} \quad B = -\frac{1 - t + 3t^2 - t^3}{1 + 3t - 3t^2 + t^3}.$$

Therefore,

$$F_{R_3}(t) = F_{\eta_2}(t)F_{\eta_3}(t)F_{\eta_3}(t) = \frac{2t^3}{1 + 3t - 3t^2 + t^3}.$$

Hence, $TH_{\text{asy}}(3, \eta) \equiv 1/E[H_1] = 1.341$. □

The above results sharpened our first intuition that the throughput may not always be monotonic in the line length. We find that, although the monotonicity result holds when there are four i.i.d. exponential tasks, it is not true under alternative processing-time distributions or different numbers of tasks. We believe that this is due to the combined effect of two factors: processing-time uncertainty and line imbalance. When we combine multiple tasks with random task times, we benefit from risk pooling (reduced coefficient of variation of station-processing times resulting from combining tasks). Consequently, the throughput per station of shorter lines is likely to be greater. On the other hand, when we combine tasks, the resulting line may become more or less balanced (in terms of expected processing times per station). If the risk-pooling benefits outweigh any imbalance created, then shorter lines are likely to be better. We also noted that the throughput of a single-station line was the highest in all the cases we considered.

Acknowledgments
We thank Kevin Angelone and the management team at Ford-Visteon Corporate and Lamosa, Mexico for introducing us to this problem and for their help in our field study. We thank the editors and two referees for several comments that helped us to improve this paper.

References


Kevin Angelone, Visteon Corporation, Manufacturing Planning, writes: “In Spring 1999, a team from GSIA, Carnegie Mellon University, headed by Professors Kekre and Rao, visited Visteon’s corporate headquarters where we explained the production-throughput and demand-backlog problems that we were facing in our rack & pinion and other reman lines. The team visited the Lamosa operations and conducted an exhaustive study of this problem over a five-month period. The team members familiarized themselves with our zero-inventory lean manufacturing environment and collected primary processing time data via time studies. At the end of the project, the results were reported to our senior management at Visteon, Dearborn. The findings included alternative approaches to capacity expansion and demand management strategies.

“I am happy to report that the ideas relating to reconfiguring the production line that were made in the project report and presentation were well received by the management at Visteon. We were provided with methods to increase production capacity merely by retraining workers and reconfiguring the line. It was shown using simulation models that the proposed changes would increase throughput by as much as 30%. A framework was presented..."
to identify the product and process characteristics for which we could expect greater benefits from shortening our lines through worker cross-training (without increasing the workforce size) or through dynamic line balancing. The results of this study have been incorporated in the changes that have been made at the Lamosa Operations. Visteon’s decision-making process for production throughput optimization has significantly been enhanced.

“I believe the results of this production reconfiguration study are well captured and summarized in the paper. This paper builds on the Lamosa project and provides practitioners with a framework for considering line-configuration choices. The field study, in our opinion, was very valuable in bringing together academics and practitioners to jointly formulate and implement models to enhance manufacturing performance at our reman operations.”