Reducing WIP in Gear Parts Manufacturing Using Queueing Networks

Volker Dörrsam, Kai Furmans and Markus Greiling*
Institut für Fördertechnik (IFK)
Universität Karlsruhe
Hertzstr. 16, D-76187 Karlsruhe
e-mail: [volker.doerrsam|kai.furmans|markus.greiling]@mach.uni-karlsruhe.de
fax: ++49/721 75 83 87

May 2, 1996

Abstract

A gear part manufacturing is modelled as a queueing network to analyze and optimize overall production performance. First a brief description of the problem is given. The second section deals with the modelling of the manufacturing system as queueing system and especially with the connection of the lotsize and the utilization respectively the coefficient of variation of the modelled system. Using a optimization approach for the variation of the lotsizes it is shown that the WIP and the sojourn time can be reduced significantly.

1 Problem Description

A major issue in the production of gears boxes for trucks is the manufacturing of the parts required for the assembly. Besides the subassembly for gear shifting and the casing the majority of parts is needed for the transmission itself, consisting of gear wheels and shafts. The manufacturing process requires several steps, which could be grouped in three sections relative to the hardening process: processing of unhardened parts, hardening and finishing of hardened parts. In the study presented we concentrated on the first section, which involves the manufacturing steps lathing, drilling, milling and gear slotting as well as washing and inspection steps.

Due to the heavy loads and torque transmitted by these parts a high quality has to be achieved which in turn requires expensive equipment for manufacturing. In combination with limited demand for truck gear boxes as well as a higher variety of available types which result in smaller annual production quantities, the manufacturing of gear parts is very often organized as a job shop with sometimes complex part routings.

The usage of one machine (or one group of machines) to produce several parts leads very often to considerable setup times, therefore the necessity arises to determine appropriate manufacturing lot-sizes for each part. The time and cost required for setups are most important with those machines, which are used to perform the very first and the last manufacturing steps. Therefore it is desirable to keep the parts that have been started in one lot together and move and process them with one setup on each machine where the lot is processed.

The previously applied lot sizing rules tended to generate some lots with very long processing times. Very often these lots would block other jobs from being processed on a particular machine, if no interruption was allowed. Due to long sojourn times it happened frequently that these other parts were needed urgently, thus forcing the interruption of the currently processed job. This happened more frequently on the machines which are typically performing the last production steps, thus leading to higher setup frequencies than planned. Unfortunately considerable setup times are needed on almost all machines, the unplanned splitting of lots into individual production lots therefore leads to a much higher loss of capacity than predicted.

After conducting a simulation study in the first step, it was concluded, that one reason for long sojourn times which in turn lead to a high probability of unplanned setups lies in the currently used lot sizing algorithm.

2 A Queueing Network Model of the Manufacturing System

Queueing network models have successfully been used to study manufacturing systems, especially the

*The work presented was part of the SFB 346 *Computer Integrated Design and Manufacturing of Parts*, financed by DFG

effects of various products competing for the same resources (see [1]).

The manufacturing system was at first analyzed with the simulation tool *DELPHI* [2], which is especially suited to support queueing network models of manufacturing systems. A few simplifications had to be made in order to provide a model that could be simulated by using DELPHI. For details about the modelling process see [3].

The simulation results showed that the sojourn times of the products covered a wide range between one or two days and several weeks. To gain an understanding of the system, a simplified queueing network model of the manufacturing unit was made, which allowed an analysis of the main factors influencing sojourn times and WIP.

The machines were grouped in disjunct sets, each set representing machines which are capable of performing the same tasks. Each group of machines is modelled as a queueing system i, where the number of machines in the group is equal to the number of servers m_i in queueing system i.

The lots of a specific part of type j are represented by a jobtype j, which requires a processing time $t_{j,i}$ on machine group i. The number of parts of type j in a lot is denoted by l_j and assumed to be constant over all process steps.

The number of operations that is performed on parttype j by a group of machines i shall be denoted by $f_{j,i}$. Typically without having scrap or alternative routings $f_{j,i}$ equals 1. For each product a demand D_j has to be covered. When product j is produced in lots of size l_j, the average starting rate for the lots of type j is $\lambda_j^{start} = D_j/l_j$. Due to practical constraints both l_j and D_j are defined:

$$l_j \geq 1, D_j \geq 1 \tag{1}$$

Combining with $f_{j,i}$ the actual arrival rate of jobs of type j at machine i is computed by:

$$\lambda_{j,i} = f_{j,i} \cdot \lambda_j^{start} \tag{2}$$

The total arrival rate of jobs at queueing system i is computed as sum over all jobtypes.

$$\lambda_i = \sum_j \lambda_{j,i} \tag{3}$$

The proportion of arriving jobs at queueing system i that is of type j, $p_{j,i}$ is defined as:

$$p_{j,i} = \frac{\lambda_{j,i}}{\lambda_i} \tag{4}$$

It is known, that the two most important factors on the beforementioned performance measures are the utilization ρ_i and the squared coefficient of variation of the service times c_i^2 at queueing system i. By assuming exponentially distributed interarrival times of the jobs at the machines, we could use an $M|G|1$-queueing system, to study the effects of lotsizing decisions on sojourn time and average number of jobs and parts waiting to be processed.

To compute these performance measures, in a first step the average service time at queueing system i has to be calculated.

$$\bar{t_i} = \sum_j p_{j,i} t_{j,i} \tag{5}$$

Next is the variance $Var(t_i)$ of the service times at queueing system i,

$$Var(t_i) = \sum_j p_{j,i}(t_{j,i} - \bar{t_i})^2 \tag{6}$$

which combined with the average service time $\bar{t_i}$ yields the squared coefficient of variation (scv):

$$c_i^2 = \frac{Var(t_{j,i})}{\bar{t_i}^2} \tag{7}$$

J_i indicates the number of different jobs at queueing system i. It is assumed, that lotsizes for jobtypes $j = 1, \ldots, J-1$ are fixed, resulting in a preliminary average service time $\bar{t_i^*}$ and Variance $Var(t_{j,i})^*$. The sum of the arrival rates of jobtypes $j = 1, \ldots, J-1$ is λ_i^* the combined demand D^* and the average number of operations on i, f_i^*. We now discuss the effects which result from determining l_J. The expressions 5 through 6 are converted to reflect the influence of choosing the lotsize for product J.

$$\bar{t_i} = p_{J,i} \cdot t_{J,i} + (1 - p_{J,i})\bar{t_i^*} \tag{8}$$

$$Var(t_i) = p_{J,i}(t_{J,i} - \bar{t_i})^2 + (1 - p_{J,i})(\bar{t_i^*} - \bar{t_i}) \tag{9}$$

The utilization ρ_i^* of queueing system i by jobtypes $j = 1, \ldots, J-1$ is calculated from:

$$\rho_i^* = \lambda_i^* \cdot \bar{t_i^*} \tag{10}$$

When $t_{J,i}^{part}$ denotes the processing time for one part of type J on a machine of group i and $t_{J,i}^{setup}$ is the average setup time required for product J on a machine of group i and the the utilization as a function of l_J can be expressed as:

$$\rho_i = \frac{t_{J,i}^{setup} \cdot D_J i \cdot f_{J,i}}{l_J} + D_J \cdot f_{J,i} \cdot t_{J,i}^{part} + \rho_i^* \tag{11}$$

The effects of lotsize variations of product J on the utilization are simple to investigate, because the first derivative is always smaller than zero.

$$\frac{\delta \rho_i}{\delta l_J} = -\frac{t_{J,i}^{setup} \cdot D_J}{l_J^2} < 0 \quad \text{for} \quad t_J^{setup}, D_J > 0 \tag{12}$$

The second derivative

$$\frac{\delta \rho_i}{\delta^2 l_J} = 2\frac{t_{J,i}^{setup} \cdot D_J}{l_J^3} \tag{13}$$

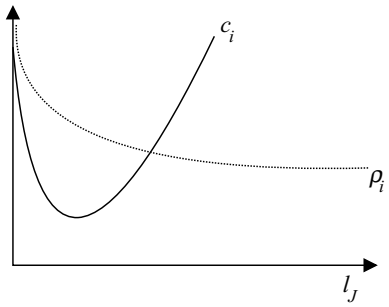

Figure 1: Effects of lotsize variations of product J on utilization $\rho_i(l_J)$ and scv $c_i(l_J)$

is always greater zero, demonstrating that $\rho_i(l_J)$ is a convex function.

The utilization decreases with increased lotsizes, because fewer setups are necessary, when fewer jobs have to be processed. The effects which can be achieved by increasing the lotsize are limited by the lower bound of the utilization of queueing system i (see 1):

$$\lim_{l_J \to \infty} \rho_i = t_{J,i}^{part} \cdot D_J \cdot f_{J,i} + \rho_i^* \tag{14}$$

When the lotsize of product J is decreased, the utilization ρ increases, approaching:

$$\rho_i(1) = (t_{J,i}^{setup} + t_{J,i}^{part}) \cdot D_J \cdot f_{J,i} + \rho_i^* \tag{15}$$

The squared coefficient of variation c_i^2 can also be expressed as a function of l_J.

$$c_i^2 = \frac{(t_{J,i}^{setup} + l_J \cdot t_{J,i}^{part} - t_i^*)^2 D_J f_{J,i} + D_i^* \cdot f_i^*}{\frac{D_J \cdot t_{J,i}^{setup_i} \cdot f_{J,i}}{l_J} + D_J \cdot t_{J,i}^{part} \cdot f_{J,i} + \lambda_i^* t_i^*} \tag{16}$$

It is intuitively clear and shown in Figure (1), that the scv reaches a minimum, when $t_{J,i}^{setup} + l_J \cdot t_{J,i}^{part}$ equals t_i^*. This can be shown by setting the first derivative $\delta c_i / \delta l_J$ equal to zero and solving for l_J. c_i approaches infinity for increasing values of l_J and approaches

$$\lim_{l_J \to 1} c_i^2 = \frac{(t_i^* - (t_{J,i}^{setup} + t_{J,i}^{part}))^2 \cdot D^* D_J}{D_J \cdot t_{J,i}^{part} D_j + \lambda_i^* \cdot t_i^*} \tag{17}$$

Therefore c_i^2 is also a convex function.

To show the effect of lotsizing variations on the average number of jobs in the queueing system, we use as already indicated a $M|G|1$-queueing system. The average number of jobs N_i in queueing system i is computed from

$$N_i = \frac{\rho_i}{1 - \rho_i} \cdot \frac{1 + c_i^2}{2} \text{ for } \rho_i < 1. \tag{18}$$

The sum of the two convex functions $\rho_i/(1-\rho_i)$ and $(1+c_i^2)/2$ is also a convex function.

By using Little's Law, and subtracting the average number of jobs in the service station the average waiting time t_i^w is obtained.

For multiple-server queueing systems, the well known approximation

$$t_w^{M|G|c} = t_w^{M|M|c} \frac{(1 + c^2)}{2} \tag{19}$$

is used. Plotting $t_i^w(l_J)$ yields figure (2).

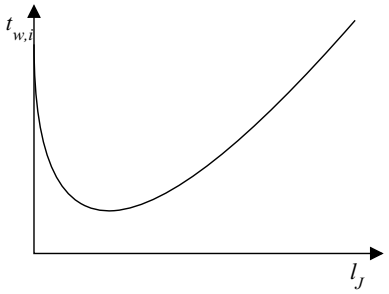

Figure 2: Expected waiting time as function of lotsize of product J

The queueing system model that has been used so far assumes that the interarrival times are exponentially distributed. This assumption is in practice not met, especially with those queueing systems which model those groups of machines which are used for the first operations on a part. Shop floor control systems usually control the starting process of the jobs in a way, that leads to a more uniform arrival process.

Though each individual part has a deterministic routing along the process flow the aggregated routing found in the manufacturing is characterized by several split and merging operations between the different production steps. Therefore the coefficients of variations of the arrival processes at the queueing systems tend to be close to 1 (see [1]). So the arrival processes at the queueing systems were modelled as Markov-Processes, which can be seen as an upper bound of the real arrival process. In addition this methodology results in a good performance of the optimization steps (see next section).

Modelling and optimizing not only single queueing systems but networks of queueing systems there are trade-off effects to be expected in form of a reduced overall variance:

Since we have to assume that the modelled system is stable and has finite capacities, the production planning has to take into account that lots with large variances of service time are to be compensated with lots with lesser variances. This means that service times of lots are not independent and results typically in negative covariances between these service times.

Since the overall variance is defined as

$$Var(\sum_i t_i) = \sum_i Var(t_i) + 2 \cdot \sum_{i<j} Cov(t_i, t_j) \tag{20}$$

we have to expect, that the overall variance is less than the sum of variances. Regarding that the proposed optimization scheme is iterative, and interim results are the basis of the next iteration, one can see, that it is necessary to calculate the network of queueing systems, and not only optimize a single queueing system (e.g. a bottleneck system).

3 Defining the optimization Problem

Using the above described simple model, the simulation results could be analyzed. As a reason for the relatively high sojourn times, squared coefficients of variations of the service which were greater than 1.0 were identified as a possible cause.

The reason for these high scvs lies in the currently used method for the lot size determination. A variation of the economic order quantity model (EOQ) of Harris, Wilson, Andler was previously used to balance expected setup costs with the cost of inventory of finished gears.

Due to the complex manufacturing processes the work-in-process (WIP) is approximately three times as large as the finished parts inventory. This part of the inventory is not part of the optimization which the EOQ performs. Therefore the usage of the simple EOQ led to lot sizes disregarding important costs induced by excess work-in-process.

Thus a a new optimization function of the form

$$\begin{aligned} \text{min:} \quad & \sum \text{Setup Costs } C_S \\ + \quad & \sum \text{Inventory Cost } C_I \\ + \quad & \sum \text{WIP Cost} C_W \end{aligned} \tag{21}$$

was created, which enriches the currently used model by taking into account the effects of lotsizing on the WIP. Modelling the job shop as a network instead of a single queue enables to regard interdependant effects between different machines and parttypes.

There is no closed form solution which could be used to compute the WIP, depending on the actual lot sizing decision. Thus an approximation based on the above described queueing network model was used to evaluate the influence of lotsizing decisions on the WIP. It is known, that overall system performance measures, like total number of jobs or average sojourn times, could be reasonably well approximated by computing the performance measures for queueing network models, even when not all underlying assumptions for the queueing network model are met (see [5]).

The cost function elements for finished goods inventory and setup costs are calculated as in the EOQ model.

$$C_S = \sum_i D_j/l_j \cdot \text{setup cost at machine } i \tag{22}$$

$$C_I = l_j \cdot \text{inventory cost per part} \tag{23}$$

$$C_W = \sum_i N_i \left(\sum_j p_{j,i} \cdot \text{WIP cost of part type } j \right) \tag{24}$$

The optimization problem being nonlinear, a gradient search method was implemented that varied the lot sizes by following a steepest descent heuristic in order to find a lot-size vector resulting in minimal total cost. Furthermore it can be shown that the underlying objective functions are convex for lotsizes equal or larger than one, which results in a definite, valid solution.

4 Application

The optimization scheme described above was implemented as a tool designed to support a group of planners for analyzing and optimizing various production areas of the large production plant. Therefore it needed to be embedded into a data-retrieving and simulation framework. To simplify file-interfaces all the input data, being a subset of the simulators data, were read directly from the simulation model description file (in ASCII format). In addition to this the optimized lotsize vector $\vec{l}$ was used to update the simulation description file.

The initial problem being analyzed, simulated and optimized was a job shop characterized by the following items:

- total number of products being manufactured: 152
- number of individual machines: 66
- avg. number of routing steps: 4.8
- max. number of routing steps: 10
- rate of products, visiting only one spezialized tool in the job shop: 35%

Some characteristics could not be regarded in the optimization model. Due to this fact a simulation tool was needed to support the modelling of the job shop. These elements are e.g.:

overlapping manufacturing The lots are produced in an overlapping manner, which results in the situation that some parts of a lot being produced on machine A, while others are already processed on the subsequently following machine B.

queueing disciplines Scheduling is done by human operators. The disciplines at different stages

of the production vary heavily. Some typical strategies are FIFO, Closest-To-Completion, Least-Slack, Dynamical-Priorization, Avoiding-Setups.

product families Some products form a "product family". Within theses families no or almost no extra setup is needed, on the other hand the switch from one family to another leads to significant setup times.

tool replacement With some machines idle time has to be taken into account caused by the time needed for replacing tools in order to produce quality on a high level.

The six steps needed – data-collection, analysis, modelling, simulation, optimization and validation – were done within approximately two days, so this optimization of a complex system could be achieved in reasonable time.

Performing the optimization for the above mentioned problem took approximately 20 minutes on a workstation and resulted in a vector $\vec{l}$ of proposed lotsizes.

The input data for the optimization were derived from the simulation model which had been used to perform the initial analysis. As the optimization model is a very much simplified model of the manufacturing system, it is not guaranteed that the solution which has been found is really an improvement over the current situation. Therefore a final simulation run was performed to ensure that the resulting lot sizes are leading to an improved situation.

The model has been further validated by comparing it with manufacturing data as well as with the experiences of the people running the job shop. This comparison was so promising that the next job shop area was also analyzed and optimized. Here the actual demand estimations were the input for the optimization. The results of the optimization were used to determine the lotsizes for that area. Figure (3) shows the effect for the WIP.

Besides the reduction of WIP by roughly 40% several more beneficial effects arose:

- Production disturbance was reduced thus implying better production scheduling;
- Lowered variance of the sojourn time helped management to predict real sojourn time better. This leads to a reduction of WIP in the *following* process-steps;
- Overall setup-rate levelled off at the same level as before. This effect is surprising, because it was expected that the setup rate would increase slightly. One possible reason is that having smaller sojourn times the splitting of lots are no longe required. While setups on first stage machines increase, later machines achieve better setup-rates than before;
- Reduced production space helped streamlining parts flow. No search operations for lost (sic!) products were needed any more.

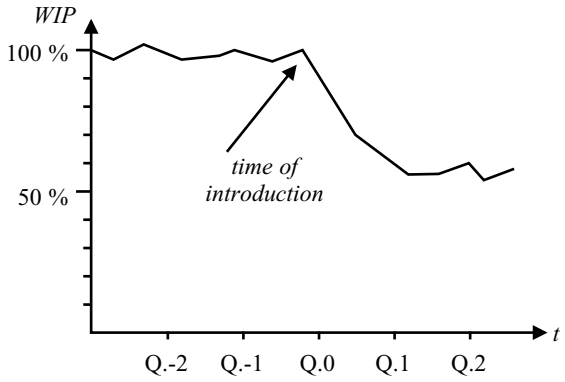

Figure 3: Work in Process before and after lotsize optimization

References

[1] J. A. Buzacott, J. G. Shantikumar 1993: "Stochastic models of manufacturing systems", Prentice Hall, Englewood Cliffs

[2] F. Chance 1994: "Delphi - Integrated Capacity Analysis and Factory Simulation", Documentation of Delphi, Version 9.7

[3] V. Doerrsam, K. Furmans 1995: "Application Case Study of a Queueing Network Simulation Tool for Analyzing and Optimizing a Manufacturing System", Proceedings of the 1995 European Simulation Symposium, Erlangen

[4] D. Connors, G. Feigin and D. Yao 1994: "A Queueing Network Model for Semiconductor Manufacturing", Submitted to IEEE Transactions on Semiconductor Manufacturing

[5] R. Suri 1983: "Robustness of Queuing Network Formulas", Journal of the ACM, Vol. 30, No. 3, July 1983