Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

SIMULATION BASED EVALUATION OF THE WORKLOAD CONTROL CONCEPT FOR A COMPANY OF THE AUTOMOBILE INDUSTRY

Patrick Kirchhof
Nicolas Meseth
Thomas Witte

Department of Operations Management and Information Systems
University of Osnabrück
Katharinenstraße 3, Osnabrück , 49069, GERMANY

ABSTRACT

This paper describes a simulation study conducted for a company of the German automobile supply industry facing the need to improve delivery reliability. The intention of the study was to evaluate whether Workload Control (WLC) is applicable as production control policy for this company and whether improvement can be expected. Therefore the regarded shop floor was modeled being organized as a WLC production system. Both, the structural and quantitative model components of the developed simulation model are explained in depth. Furthermore it is shown how inherent parameters of the WLC concept can be set using the simulation model in a practical environment. As due date compliance is the primary concern of the company, the performance of four simple priority dispatching rules is analyzed with regard to delivery reliability. As a result it is shown that WLC is applicable in the given situation and that performance enhancements can be expected.

1 INTRODUCTION

Simulation of production systems proves particularly advantageous when strategic decisions such as the selection of an appropriate production control policy have to be taken. Information about a complex real system can be derived from a simulation model of the real system and support decision making (Seila, Ceric, and Tadikamalla 2003).

This simulation study was conducted in cooperation with a company of the German automobile supply industry that perceived the need to improve poor delivery reliability and reduce delivery times. Misled by its promising characteristics, it implemented a Kanban production system. However, its applicability was not verified sufficiently prior to the implementation and the outcome was disappointing. Delivery reliability fell sharply and delivery times became hardly predictable due to their high dispersion. A subsequent investigation revealed that the company did not comply with most of the requirements needed to apply Kanban successfully and the previous system was reestablished at high costs. As reported by Henrich (2004), this is not an isolated incident and the selection of a production control policy is frequently based on intuitive reasoning instead of the evaluation of the company's characteristics.

This paper describes the simulation study later on conducted for the mentioned company in order to analyze the applicability of Workload Control (WLC) as production control policy. Furthermore, it is pointed out how WLC can be parameterized appropriately in a practical WLC environment using simulation. The intention is to report about the study and its results and to give insight to the methodology.

In the following section WLC is presented and its basic principles are explained briefly as the focus of this paper lies on the conducted study. Hereafter, the experimental design is defined and the simulation model is elaborated. It is developed using the simulation software Arena by Rockwell. Section 4 illustrates the parameterization of the system in order to fully specify the simulation model. The performance of WLC is then assessed in section 5 employing the simulation model. The evaluation contains a sensitivity analysis that examines the system's behavior for varying input data. Finally, the most relevant aspects of this study are summarized in the conclusion.

2 WORKLOAD CONTROL

2.1 General Principles

WLC is a load oriented production control policy intended to establish short and precisely predictable lead times in order to improve delivery reliability. It is based on the relation of work in process (WIP) and throughput time which was first expressed by Little (1961) and is today known as Little's Law. It implies that mean throughput times can be decreased by reducing the mean WIP. This relation holds for any steady state system regardless of variations in input or output sequence. Thus, the philosophy of the policy is to create short, stable and predictable queues in the production

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

process in order to minimize throughput times. The main instrument to control WIP is the release decision, transferring an order to the set of orders admissible for production. The criterion for this decision is the level of workload at each work station which the order has to pass according to its routing sequence. Workload levels are usually measured in units of processing time and are compared to workload norms. These are assigned to the work stations and serve as a reference value in the release decision (Lödding 2005).

Different methods to compute workload levels are presented in literature. All approaches distinguish between direct load and indirect load. Direct load comprises all orders queueing at the considered station. All upstream orders that prospectively pass the considered station according to their routing but are queued or processed at a previous work station are referred to as indirect load. The most straightforward method introduced by Kingsman, Tatsiopoulos, and Hendry (1989) is adopted in this work. It determines the *aggregate load* by adding direct and indirect load of a work station. The approach developed and presented by Bechte (1984) at the University of Hannover employs a method called *load conversion* to estimate upcoming direct loads. The sum of direct and estimated load is called *converted load* and subject to the norm. Oosterman (2000) recommends that aggregate load should be corrected according to the relevant routing of the order and proposes the *adjusted aggregate load*. An extensive review on this matter is e.g. given by Stevenson and Hendry (2006), Breithaupt (2002).

Land (1996) introduced the *Decision Framework* which describes the decision process within the WLC concept. Three decision moments are distinguished: entry, release and dispatching. Decisions in the entry level are used to control the medium term level of load in the whole production system. The objective of the release decision is to control the workload on the shop floor, whereas the dispatching decision regulates the progress of individual jobs at the work stations. Precise production schedules and their exact realization are not necessary and do not exist in WLC, solely ordinary priority rules control the order flow at the work stations (Kingsman 2000). A detailed description of the WLC concept and the decision process is presented by Land (2004).

2.2 Applicability

WLC is especially designed and well suited for Job Shop Production (JSP). JSP environments are characterized by a high variability of batch size, routing, operating times and products that are manufactured. Routing can vary in regard to length and sequence. Workstations are grouped functionally considering their capability, whereas parts are transported batchwise according to their designated routing. In its immaculate form - the pure JSP - any station can be predecessor or successor in a routing sequence. Yet these assumptions are very stylized and do not pertain in most real life job shops. They do usually reveal a predominant flow direction since certain machines typically perform preparatory operations and others perform terminatory ones. This is best described by the theoretical general flow shop (GFS) which is characterized by deviations in terms of routing length but not routing direction (Oosterman 2000). A completely directed flow resembles a flow shop which can not be controlled under the WLC regime. Further criteria for the application of WLC and their influence on the performance e.g. the order structure, due date requirements or processing times are presented by Henrich (2004).

3 EXPERIMENTAL DESIGN

3.1 Shop Floor Characteristics

In the observed company, six groups of items are produced on fourteen groups of machines. Routing and operating times are dependant on the group of item. Routing length can vary from four to seven operations and no reentrant loops are observed but a dominant flow direction clearly prevails. The order flow is presented graphically in figure 1. The ellipses represent groups of machines, the colored arrows indicate the order routing of the respective group of items.

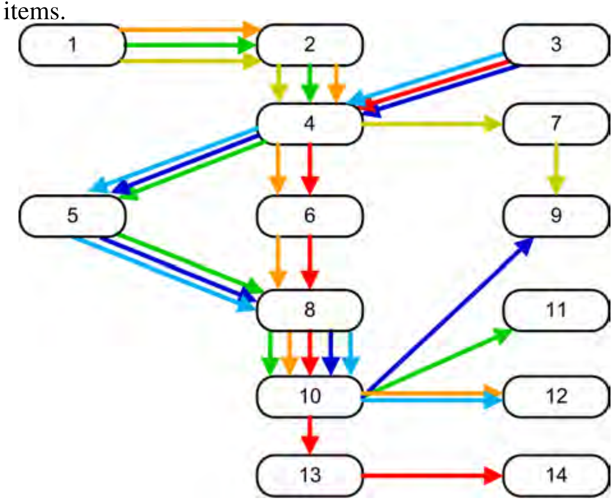

Figure 1: Graphical representation of the order flow in the considered shop floor.

3.2 Model Overview

As stated the simulation model was developed using the simulation software Arena. Submodels are designed and named in accordance with the decision framework discussed above. In the basic model, the priority criterion at the release stage is the due date of the order itself, while the workstations apply the first-come-first-serve priority rule (FCFS), a setup

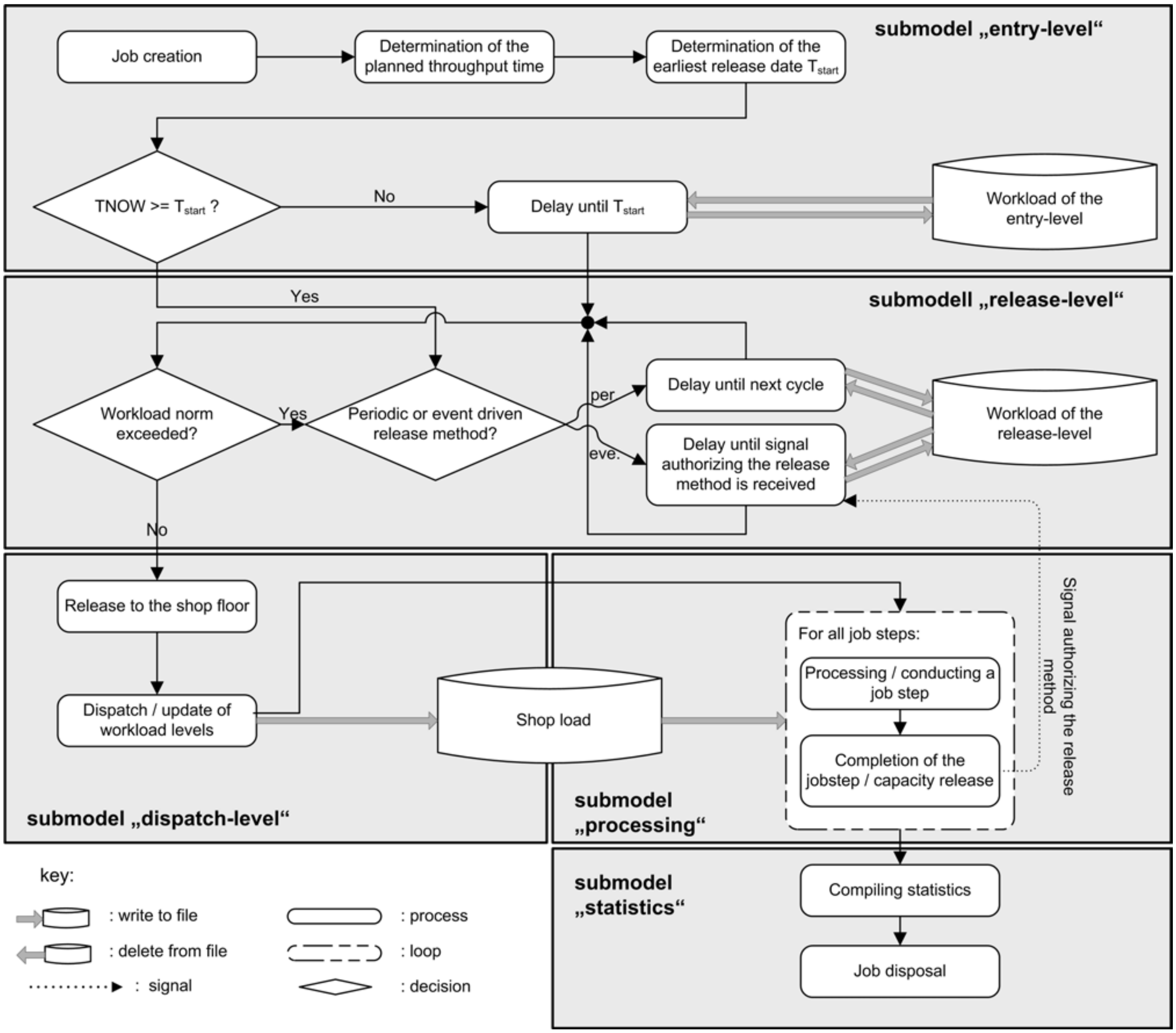

Figure 2: Overview of the simulation model.

that is commonly used. Figure 2 shows the execution of the simulation model on a very high level of abstraction. The following subsections describe excerpts of the model in greater depth.

3.3 Structural Model

Orders are generated with individual inter-arrival times for each group of items respectively to their determined statistical distribution. Relevant attributes such as due date or order quantity are assigned. In order to explore the influence of urgent orders (i.e. orders that are received being already behind schedule) on the system's performance these are created as well. The percentage of urgent orders and the degree of lateness are customizable.

The planned release date is then calculated for each job by scheduling back from the due date as described in the previous section. If the planned release date already lies in the past, the order is threatened to be late and immediately released onto the shop floor. This might be the case for orders with very large order quantities or urgent orders. Yet, if configured properly most of the orders will be held back until their planned release date is reached. With a `Delay` block the individual waiting time is dynamically assigned to the order.

Both, the periodic and the event driven release method were incorporated into the model and can be activated as desired. The method invoked is the same in both cases, however the difference lies in the mode in which it is accomplished. The periodic method is time based and the release function is called in predefined intervals in simulation time. All entities that reach their individual planned release date within the interval are queued and marked as release candidates for evaluation in the following cycle. This procedure disrupts the inflow of orders before it releases multiple jobs at a time. As a consequence, the levels of WIP vary following a typical saw tooth pattern. This effect becomes more distinct the lower the mean WIP levels are. Release period length and mean WIP must therefore be aligned in an adequate proportion (Stevenson 2005).

The event driven method employs internal messages as authorization in order to release jobs on the occurrence of certain events. In Arena, these messages are implemented with `Signal` blocks which are configured to submit a signal whenever capacities are cleared, i.e. when an operation is completed and the relevant lot is transported to the subsequent work station. `Wait` blocks holding back the release candidates receive those signals and allow for the release function to be performed on all orders located in the queue. Hereby the inflow of orders is continuous which stabilizes WIP levels. In turn it requires more accurate feedback information from the shop floor.

The release function itself sequentially validates for all queued orders whether releasing it onto the shop floor would cause any workload norm to be exceeded. If this is not the case, the order is passed on to the dispatch level. Otherwise the order is put back into the queue for the next release procedure. As simulation time proceeds, its priority and therefore its likeliness of being released will increase due to the diminishing slack.

Direct and indirect load of the work centers are updated according to the respective routing when orders reach the dispatch level. Hereafter the order is placed in the first queue of its relevant routing. The production process was modeled following the station concept available in Arena. Thereby the explicit modeling of numerous work stations and repeatedly similar program routines can be avoided. The `Sequence` element allows for the definition of routings and their respective operating times. Complex production systems can be modeled comprehensibly by specifying a generally valid structure (Witte, Claus, and Helling 1994). As an operation is completed, the workload of the respective work station is reduced and changes are stored. In case of the event driven release method a signal authorizing the release procedure is sent.

The simulation model contains a submodel that is intended to collect statistical data during the simulations run. It compiles statistics of e.g. lead times, delivery times, delivery reliability, utilization levels and the WIP as performance measures.

3.4 Quantitative Model

The correct representation of all logical aspects of the simulation model is subject to the structural model as opposed to the quantitative model, whose subject-matter it is to determine all quantitative model components. Processing times and routings were adopted directly from the company's enterprize resource planning system (ERP). The statistic distributions of the inter-arrival times, order quantities and the percentage of urgent orders were derived from historical data as proposed by Witte, Claus, and Helling (1994). At first, the kind of statistical distribution is estimated by means of histograms. A histogram is a graphical approximation of the density function's graph and hence it contains a reference to the underlying distribution. Then the respective parameters are estimated which yields an assumed probability distribution. The relevance of the statistical model can be assessed objectively by goodness of fit tests such as the Kolmogorov-Smirnov test or the chi-square test.

For the arrival process, an exponential distribution *Exp*(λ) was assumed. The reciprocal of the mean inter-arrival times was used as the point estimator for the arrival rate λ and confirmed by the chi-square test. Order quantities are designed to be triangularly distributed as shown in table 1. Due date allowance follows a normal distribution *Norm(30;6)* for all groups of items. The percentage of urgent orders and their time for delivery could not be retrieved from the company's data. Interviews with domain experts were conducted and led to the assumption that 5 % of the orders come in urgent and that their time for delivery can be best described as triangular *Tria(3;4;5)* distributed.

Table 1: The derived probability distributions used in the simulation model.

item group	distribution of the inter-arrival times	distribution of the order quantities
1	Exp(2,31)	Tria(2023 ; 2693 ; 3760)
2	Exp(1,23)	Tria(1198 ; 2642 ; 4120)
3	Exp(2,18)	Tria(1000 ; 2515 ; 4965)
4	Exp(1,89)	Tria(1500 ; 2250 ; 9000)
5	Exp(1,92)	Tria(515 ; 1875 ; 3800)
6	Exp(2,83)	Tria(741 ; 1509; 2000)

4 DETERMINATION OF PARAMETER VALUES

Its robustness and its simplicity account for the major advantages of WLC. Only a few and simple rules make up a sophisticated and highly efficient production control policy. Yet, the adjustment of the relevant parameters of the system is a very crucial task as reported by Land (1996), Land (2006). The system's performance is highly dependant on these parameters and finding an appropriate setup is delicate, especially for parameters that lack an analytic procedure of determination. A systematic variation of the parameters in the simulation model can be performed in order to find a sensible configuration.

4.1 Workload Norms

High workload norms cause long lead times with higher dispersion, whereas tight norms decrease utilization levels and prevent a working load-balancing. Analytical approaches to determine the values of the workload norms are sparse. Two different suggestions are presented in (Harkose, Kingsman, and Worthington 2004, Nyhuis and Wiendahl 1999). In this

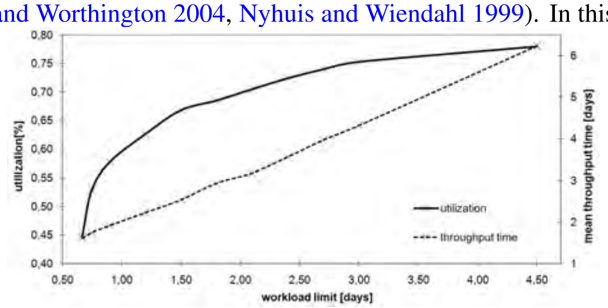

Figure 3: Collective determination of the workload limits.

study, at first parameter values were analyzed collectively. Limits were systematically increased starting with a load limit of only two shifts length (16 hours) up to a limit of 4.5 days. They are measured in time units according to the planned operating times. Considering the results given in figure 3 the workload was limited to 2.25 days of work for each work station. Individual adjustments were then made for potential bottlenecks. Endorsing the proposal of Oosterman (2000) workload limits could be adjusted according to their relative position in the order flow direction. This is relevant for a GFS since work stations at the end of the production process experience higher levels of indirect load than stations at the beginning of the process.

4.2 Release Period Length and Release Method

The longer the release period length is, the more disturbances are caused in the influent order stream. Orders are delivered batch wise and batch size increases in correlation to the release period length. In turn, delivery reliability decreases at longer intervals. This behavior is expected, yet it is observed to be intensive for the given data. As a normalized measure of dispersion the variation coefficient of the range of coverage reveals the deviations of the WIP provoked by the release period. Even a release period length of a single day, which is commonly used, underperforms significantly. Owing to the much higher performance, the event driven release method, which resembles a marginally short release period length, is adapted hereafter.

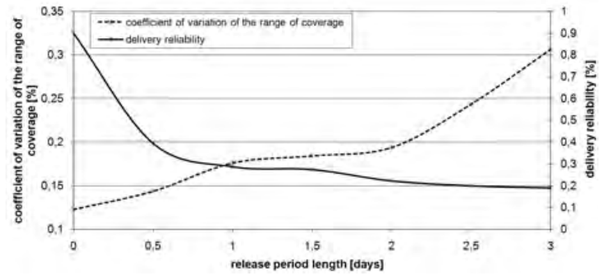

Figure 4: Analysis of the release period length.

4.3 Priority Dispatching

It is generally agreed upon to use the due date of the orders as the priority criterion in the release decision. However, the applicable priority rule of the dispatching decision is controversial, although FCFS is most often used. Its performance was compared to that of the shortest slack per operation rule (S/OPN), the critical ratio rule (CR) and the shortest processing time rule (SPT). Both, S/OPN and CR are intended to reduce the dispersion of lateness and belong to the group of due date oriented priority rules. Slack time divided by the remaining number of operations yield the priority criterion of the S/OPN rule. At this orders gain priority over time as slack decreases and are then preferably processed. CR determines the ratio of time remaining until the due date and the remaining operating times which is used as the priority index. This slows down orders that are early and accelerates late ones (Vollmann, Berry, and Whybark 1997). According to the SPT rule, the order with the shortest processing time on the workstation is considered first for dispatch which generally leads to high utilization and shorter mean throughput times at the cost of higher dispersion of throughput times and weaker delivery reliability (Hopp and Spearman 2001).

Figure 5 shows the comparison of these four rules with regard to utilization and figure 6 with regard to delivery reliability, both for a variation of WIP. Like workload limits, the WIP levels are measured in units of time corresponding the planned operating times. Utilization is not significantly affected by the dispatching rules, even though SPT performs slightly better. With respect to delivery reliability the differences are more relevant. CR and S/OPN perform better on a wide range of WIP levels and are therefore pursued.

5 SIMULATION EXPERIMENTATION

In this section the results yielded by the fully specified and parameterized simulation model are presented. Over 90

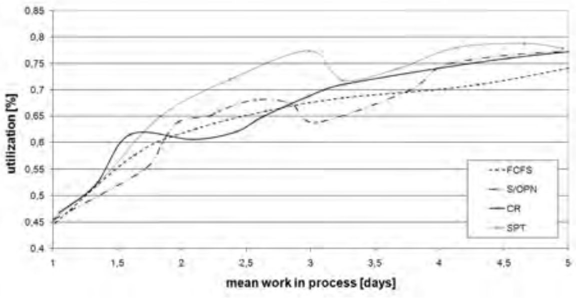

Figure 5: Comparison of priority dispatching rules with regard to utilization.

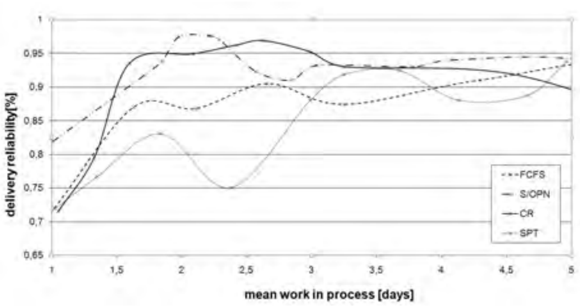

Figure 6: Comparison of priority dispatching rules with regard to delivery reliability.

simulation runs were analyzed each with a length of 300 days including a warmup period of 100 days.

5.1 Results

Mean utilization is at 70 %, with utilization of the bottleneck station at 90 %. The mean average lead time is 3.6 days with a standard deviation of 0.8. Mean processing time of an order is 1.3 days, which implies that lead time amounts to less than the triple of the pure processing time. Bearing in mind the number of operations performed on each job and the respective number of queues, this is remarkable. Delivery reliability is accordingly high with 98.8 %. This includes rush orders whose mean time for delivery is only four days, i.e. merely longer than the mean lead time. Apparently the adaption of delivery date oriented priority rules in the dispatch decision significantly facilitates the handling of rush orders. Minimal lead times can be achieved that are by far shorter than using conventional throughput oriented or flow oriented rules. Yet, many permutations must be accepted in the queues as priorities and therefore queueing sequences are constantly changing. Disturbances on the shop floor are the downside of the attained flexibility.

5.2 Sensitivity Analysis

A sensitivity analysis was conducted with the parameterized system in order to investigate the influence of changing input data on the system's behavior. Input data of the described standard system was gradually modified by +/- 30 %. As controlled variables the percentage of rush orders, order quantity, inter-arrival rate, times for delivery and machine failures were determined. Variation of the results is measured relatively to the standard model. Figure 7 presents the results concerning the delivery reliability. Expectedly the delivery reliability increases with decreasing order quantities and vice versa. Since the reference value in the basic model is a high one, delivery reliability can only increase slightly, whereas it decreases sharply for larger order quantities. For the same reason the delivery reliability remains constant on increasing inter-arrival times. In the given range the system reacts invariant to a change of the order fulfillment time as it is sufficiently long with the 5.5 fold mean processing lead time. If rush orders occur more often than in the standard configuration it causes deliver reliability to decrease significantly. The same holds for machine failures. This is plausible because the system operates on high levels of utilization that do not offer additional vacant resources. For the given structure of input data only expanding capacities could make the system more invariant to external disturbances.

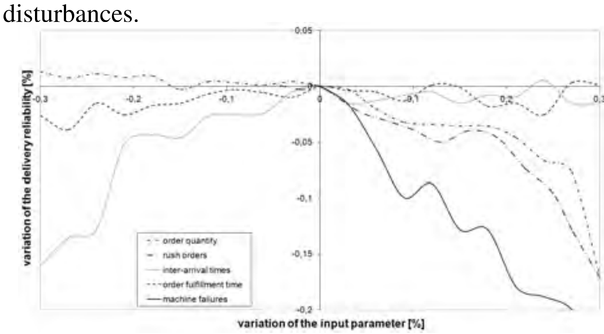

Figure 7: Sensitivity analysis of the simulation model for the relevant input data.

6 CONCLUSIONS

The simulation study presented in this paper is a good example of how simulation can facilitate a complex decision process. It shows how detailed information about the performance of a prospective production control policy can be obtained prior to its implementation. Besides it was shown how such a hypothetical system can be configured using simulation. A set of sensible parameter values can be obtained by systematic parameter variation. This is surely a pragmatic approach and far from an optimization in the mathematical sense, but not to be underestimated either as performing a systematic search contributes strongly to a deep understanding of the system's dynamics.

REFERENCES

Bechte, W. 1984. *Steuerung der durchlaufzeit durch belastungsorientierte auftragsfreigabe bei werkstattfertigung*, Volume 70 of *Fortschritt-Berichte VDI*. Düsseldorf: VDI-Verl.

Breithaupt, J.-W. 2002. The workload control concept theory and practical extensions of load oriented order release. *Production planning & control : PPC* 13 (7):625–638.

Harkose, A., B. Kingsman, and D. Worthington. 2004. Performance analysis of make-to-order manufacturing systems under different workload control regimes. *International journal of production economics* 90:169–186.

Henrich, P. 2004. Exploring applicability of the workload control concept. *International journal of production economics* 90 (2):187–198.

Hopp, W. J., and M. L. Spearman. 2001. *Factory physics*. 2nd ed. Boston: McGraw-Hill/Irwin.

Kingsman, B., I. Tatsiopoulos, and L. Hendry. 1989. A structural methodology for managing manufacturing lead times in make–to–order companies. *European Journal of Operational Research* 40 (2):196–209.

Kingsman, B. G. 2000. Modelling input-output workload control for dynamic capacity planning in production planning systems. *International journal of production economics* 68 (1):73–93.

Land, M. J. 1996. Workload control concepts in job shops a critical assessment. *International journal of production economics* 46/47:535–548.

Land, M. J. 2004. *Workload control in job shops, grasping the tap*. Ridderkerk: Labyrinth Publ.

Land, M. J. 2006. Parameters and sensitivity in workload control. *International journal of production economics* 104:625–638.

Little, J. D. C. 1961. A proof for the queuing formula: $l = \lambda \cdot w$. *Operations Research* 9:383–388.

Lödding, H. 2005. *Verfahren der fertigungssteuerung*. VDI. Berlin: Springer.

Nyhuis, P., and H.-P. Wiendahl. 1999. *Logistische kennlinien*. VDI. Berlin: Springer.

Oosterman, B. 2000. The influence of shop characteristics on workload control. *International journal of production economics* 68:107–119.

Seila, A., V. Ceric, and P. Tadikamalla. 2003. *Applied simulation modeling*. Duxbury applied series. Belmont, Calif.: Thompson.

Stevenson, M. 2005. A review of production planning and control the applicability of key concepts to the make-to-order industry. *International journal of production research* 43 (5):869–898.

Stevenson, M., and L. Hendry. 2006. Aggregate load-oriented workload control a review and a re-classification of a key approach. *International journal of production economics* 104 (2):676–693.

Vollmann, T. E., W. L. Berry, and D. C. Whybark. 1997. *Manufacturing planning and control systems*. 4th ed. New York NY: McGraw-Hill/Irwin.

Witte, T., T. Claus, and K. Helling. 1994. *Simulation von produktionssystemen mit slam*. Bonn: Addison-Wesley.

AUTHOR BIOGRAPHIES

PATRICK KIRCHHOF is a research assistant at the department of Operations Management and Information Systems at the University of Osnabrück, Germany. He received a diploma in Business Administration with a major in Information Systems in 2007. His research interests are simulation modeling and production control. His email address for these proceedings is <Patrick.Kirchhof@Uni-Osnabrueck.de>.

NICOLAS MESETH is a research assistant at the department of Operations Management and Information Systems at the University of Osnabrück, Germany. He received a M. Sc. in Information Systems in 2006. His research interest are simulation modeling and production control. His email address for these proceedings is <Nicolas.Meseth@Uni-Osnabrueck.de>.

THOMAS WITTE holds the chair of Operations Management and Information Systems at the University of Osnabrück, Germany. His email address for these proceedings is <Thomas.Witte@Uni-Osnabrueck.de>.