Reliability of Intra-Logistics Systems – Simulating Performance Availability
Steffen Schieweck, Johannes Dregger, Sascha Kaczmarek, Michael ten Hompel

Abstract—Logistics distributors face the issue of having to provide increasing service levels while being forced to reduce costs at the same time. Same-day delivery, quick order processing and rapidly growing ranges of articles are only some of the prevailing challenges. One key aspect of the performance of an intra-logistics system is how often and in which amplitude congestions and dysfunctions affect the processing operations. By gaining knowledge of the so-called ‘performance availability’ of such a system during the planning stage, oversizing and wasting can be reduced whereas planning transparency is increased. State of the art for the determination of this KPI is simulation studies. However, their structure and therefore their results may vary unforeseeably. This article proposes a concept for the establishment of ‘certified’ and hence reliable and comparable simulation models.

Keywords—Intra-logistics, performance availability, simulation, warehousing.

I. INTRODUCTION

Spatiotemporal transformation processes within a building are performed by intra-logistics systems [1]. Depending on their type and scope, intra-logistics systems can consist of highly complex structures and several subsystems (materials handling, storage area, picking area, packaging area, etc.). Examples of complex intra-logistics systems are warehousing systems, parcel distribution centers or baggage handling systems at airports.

Performance is often used as a key indicator for the qualitative assessment of an intra-logistics system. It is commonly understood as the throughput of an intra-logistics system, e.g. how many objects are transformed in space and time within a defined observation period [2].

Considering the performance of modern, complex intra-logistics systems in an isolated manner cannot be considered adequately. Due to increased competition, high demands must be met by intra-logistics systems - not only regarding performance but especially availability. From the viewing point of the operating company the intra-logistics system has to operate continuously and upcoming risks must be quantified and minimized.

The presented research is funded by the German Research Foundation (DFG) within the subproject C1 “Analytical Methods for the Determination of Performance Availability of Complex Material Flow Systems”, which is part of the project PAK 672 (HO 2403/8-1).

Steffen Schieweck is with TU Dortmund University, 44137 Dortmund, Germany (phone: +49 231 755-4832; fax: +49 231 755-4768, e-mail: steffen.schieweck@flw.mb.tu-dortmund.de). Johannes Dregger, Sascha Kaczmarek, and Michael ten Hompel are with TU Dortmund University, 44137 Dortmund, Germany (e-mail: johannes.dregger@flw mb tu-dortmund.de, sascha.kaczmarek@flw mb tu-dortmund.de, michael.tenhompel@flw mb tu-dortmund.de).

To meet these requirements the KPI of performance availability has been developed. In contrast to performance which only represents the (maximum possible) throughput, performance availability considers the degree by which the desired performance can be fulfilled over a specified period. This is in contrast to methods that solely consider availability in terms of dysfunctions (cf. [3], [4]).

Gaining knowledge during the planning stage of the implicit performance availability of an intra-logistics system is crucial for any designer. To our knowledge, currently no analytical method has been established for the prediction of performance availability. The approaches of [5] and [6] are starting points which need to be further developed and expanded. At present, simulation studies are commonly used for this step. However, as their structure and measuring methods are not specified, their results may vary unforeseeably. To reach the state of being able to create and use ‘certified’ and hence comparable simulation models we present an approach for the standardization of simulation models for the assessment of performance availability. Also, we aim to spread the approach of performance availability on an international level as it has scarcely been a topic of interest in English literature.

The article continues with a general depiction of the state of the art, literature review and measuring methods of performance availability in Chapter II. Chapter III presents the proposed modeling paradigm which is subdivided into the sections structure, data and measurement. To deepen the understanding of performance availability and its measurement a case study is conducted in Chapter IV before we draw basic conclusions in Chapter V.

II. PERFORMANCE AVAILABILITY

System functionalities and performance requirements of intra-logistics systems are stated contractually between a systems operator and vendor and must be verified after launch. For that, the VDI guideline 4486 provides standard procedures. It focusses on business objectives and their fulfillment. The concept of performance availability is herein defined as:

"The performance availability indicates the degree of fulfillment of processes agreed between contract parties (manufacturer and user) in accordance with the requirements and deadlines and in compliance with the agreed basic conditions." [7]

Based on this definition the guideline describes a method for determining performance availability during the validation procedure of the functionality of the intra-logistics system. The measurement methods are described in greater detail in

Keywords: Spatiotemporal, intra-logistics, simulation, performance availability.
the following Chapters A-E.

A. Literature Review and Research Outline

The concept of performance availability has first been defined by Wittenstein [8] in the context of customer-specific product development for the field of machine and plant engineering. Here, performance availability is described as the state in which a process is carried out properly and the required result can be completed on schedule [8]. This means that at the time of the request the necessary provision of the service resources are available, regardless of uncertainties such as fluctuations of demand or dysfunctions [8]. Performance availability thus describes the state of a system in which one or more processes can be performed such that the results meet the requirements in both time and quality, and hence a demand oriented service can be provided.

The basic idea of Wittenstein has been employed by the VDI guideline 4486 and specified in relation to the measurement of processes in intra-logistics systems. The parallels between VDI 4486 and the definition of Wittenstein are clearly visible. Decisive, however, is the concretization of the description of performance availability from a state to a quantitative, measurable size.

The VDI guideline exclusively focuses the validation procedure for functionality after the launch of the intra-logistics system. Maier [9] deals with acceptance procedures for intra-logistics systems, with a focus on the analysis of existing methods and the derivation of requirements for acceptance procedures with the help of business cases. Maier concludes that the existing procedures and existing KPIs are not or only partially suitable for the acceptance procedures of intra-logistics systems. Also, she proposes a method for the analytical prediction of performance availability [6].

B. Measurement Methods for Performance Availability

Unlike Wittenstein, the VDI guideline 4486 defines performance availability as a quantifiable variable which reflects a degree of satisfaction. The measurement of performance availability can be done using the measurement of timeliness (1) or waiting time (2). The respective degree of performance is given by the number of on-time material flow objects compared to the quantity of all of the arriving objects during the observation period or the ratio of the difference between the observation period and cumulative waiting times during the observation period (cf. [7]).

When calculating the performance availability based on timeliness \( \eta_l \), the number of timely objects are detected either directly or determined from the total number of objects \( N \) minus the number of tardy objects \( n \). (VDI 4486 uses the term “running times” instead of timeliness).

\[
\eta_l = \frac{N - n}{N} \quad (1)
\]

For the calculation of performance availability based on waiting times \( \eta_w \), the single objects are not assessed individually, but the cumulative waiting time \( T_w \) within a defined observation period \( T_B \) is recorded.

\[
\eta_w = \frac{T_x - T_w}{T_w} \quad (2)
\]

C. Parameters and Measuring Points

The determination of performance availability requires the collection of different data. These need to be determined at various points of the intra-logistics system. Typically this is the last resource involved in each transformation and where the objects leave the process.

D. Acceptance of Downstream Processes

Intra-logistics processes can also be part of a series of physical transformation processes and thus include upstream or downstream processes. The passing of objects at the corresponding interfaces depends on the operation of the linked downstream process, i.e. the process is able to provide or to handle the objects. For this reason, it is necessary to consider the operation of downstream processes or systems for the measurement of performance availability.

E. Observation Period

The observation period describes the time interval for which the parameters are collected. The performance availability is determined on the basis of this observation period. Since there are time-varying variables, the choice of the length of the observation period has influence on the resulting performance availability.

III. MODELING PARADIGM

The proposed paradigm aims to serve as a guideline and first step to the creation and measurement procedure for the definition of performance availability. Our presentation of the modeling paradigm will be divided into three sections, the modeling structure, data and performance availability measurement. It must be noted that this segmentation is undertaken only for the purpose of presentation. Certainly the creation of a simulation model requires thorough planning of the mentioned components under constant consideration of their interferences.

Fig. 1 Modeling paradigm

Within the structure section the requirements of the modeling structure and the degree of completeness of the replication of the real-world system is described. The data section covers the treatment and/or creation of input data used...
for the simulation runs. The measurement section specifies the procedure of performance availability measurement as defined by [7] and described in Chapter II. Within this context the understanding of the guideline is discussed critically.

A. Structure

The structure of a simulation model can hardly be predetermined even when being constrained to one domain, such as warehousing or intra-logistics. Nonetheless, the structure has to fulfill requirements for the upcoming measurement of performance availability.

As stated in Chapter II, the KPI performance availability refers to a whole intra-logistics system. Hence, the whole system or at least its core subsystems must be included in the model. This might include receiving, shipping and storage departments as well as unpacking, putting, picking, consolidating, packing and transportation operations and structures. It is not compulsory to include all of the aforementioned systems. However, as performance availability is designated to be part of contractual agreements it is highly recommended to include all of the subsystems which are part of the contractual documents or have significant influence on them.

Particular attention should be paid to the modeling of the transportation systems which are the connections between the subsystems. Concerning performance availability they feature special significance as they also serve as buffers between various subsystems. The performance of a warehousing system is substantially reduced if a queue of loads is large enough, so that the antecedent system cannot release its completed job to the transportation system (aka the buffer is full). Thus, the exceeding workload of one subsystem affects the other subsystem’s performance. This aspect is not covered in the conventional understanding of availability. Hence, a fundamental aspect of a model to measure performance availability is to accurately represent a transportation systems buffer capacity, its conveying speed and length.

As the measurement of performance availability (or one of its specifications) is based on the waiting time of crucial resources, all resources of concern must be included in the model. This might be employees such as pickers and packers or automated resources such as stacker cranes. Assigning states to the resources such as “idle,” “waiting for release,” “busy” and so forth simplifies the tasks of waiting time measurement significantly.

B. Data

The utilization of simulation studies implicates the availability of data. The data can either be available from a real-world use case or need to be created for the virtual testing of the designed system. For both cases, guidelines are proposed in the following chapter. Generally, [7] proposes to specify a time frame for the testing of performance availability in the real-world system. In most cases this will be a minimum amount of time in which representable data can be created. Simulation serves the advantage of being able to extend this time frame at very low cost and hence is able to create a higher degree of representativeness. The amount of data created (or available) should be assessed with respect to this fact.

Naturally, data extracted from a real-world system are preferable. Concerning the applicability of its results a simulation model is only as good as its input. Even if data are available certain fundamentals have to be accounted for. Firstly, a representable set of data has to be selected that is within the performance specifications of the considered system (see also [7]). Secondly, the data ought to pose some degree of challenge to the system and thus depict its load in peak periods. Some performance availability characteristics do not reveal if the required performance is constantly low (which means, considerably lower than the maximum performance).

If no real-world data are available, their manual creation is inevitable. Again, the performance requirements should lie close to the maximum performance of the system, but not above. For the case that the performance is specified in numbers of operations per hour the system can be challenged by creating times with low performance requirements and high peaks. Again, the peaks have to lie within a reasonable amplitude and all of the specifications. However, creating ‘peaking’ data is not preferable if the main business objective is specified as the utility of resources. Parameters to consider for the creation of data are the number of operations per hour (picking lines, orders), the overall inventory, the inventory per SKU, and the lines in proximity.

By the latter index we believe the throughput time of an order and hence to some degree the performance availability is significantly affected. For the purpose of this paper, we define the lines in proximity as the number of order-lines which access the system at the time of and within a certain amount of time (proximity) before a considered order. Note that the number of order-lines of the considered order itself is included in the lines in proximity. With \( p_{i,p} \) being the lines in proximity \( P \) of order \( i \), \( l_j \) the number of lines of order \( j \), \( T_p \) the set of accessing times within proximity \( P \) and \( t_j \) the accessing time of order \( j \) into the system the formal definition is as follows:

\[
p_{i,p} = \sum_{l_j} t_j \quad \forall i
\]

with \( T_p = \{ t_j \mid t_j \leq t_i < t_i + P \} \).

The justification of the proposed index is trivial. Testing performance availability, the system is expected to deal with congestions and queues. The length of a queue, its probability of existence and therefore the expected waiting time of an order and its comprised storage totes highly depends on the number of transportation operations which currently stress the system. The value of the proximity \( P \) has to be chosen such that the past influencing orders which might lead to waiting times are regarded appropriately. An approach to that matter is given in Chapter IV.
For the measurement of performance availability based on running times (see [7] and Chapter II) the creation of due-dates is necessary. This has to be conducted in a reasonable manner and is supposed to relate to real-world applications by a large degree.

Due-date creation has been a topic of interest for many years, publications can be found as early as 1967 (e.g. [10]). The motivation for most of the research in this area is the prediction of flow times through a job shop and therefore the creation of due-dates with low tardiness. Surveys for this area of research are given in [11]-[13], to name but a few. Generally, due-date assignment methods are subdivided into two categories, externally (exogenous) set due-dates or internally (endogenous) set due-dates [14].

It must be noted that for the purpose of this research the determination of due-dates with low tardiness might not be suitable. For our work, due-dates must be created as they appear in real-world warehouses. Hence, we consider comparatively simple rules from the literature. To fulfill the needs of our work, we also developed an additional method. Following, some rudimentary rules are described briefly. After, our own simple approach is presented.

The constant flow allowance (CON) method allows each job the same running time from its access to the systems until completion [15]. It is part of the exogenous methods as no information from the flow system is required for the determination. The amount of flow allowance might be chosen freely, however for practical consideration the average running time through the system should be held for a reference value. The random method (RAN) gives any job a random flow allowance which is uniformly distributed within a specified interval. Exemplary for an endogenous method the number of operations (NOP) defines due-dates as a function of the number of operations which have to be performed for a considered job. For this application we propose either the use of order-lines of an order or the aforementioned lines in proximity.

An order cannot be assigned to a route if the remaining time until its departure is smaller than a critical value $h$. If this is the case the order gets assigned to the next batch on the same route. By this manner we prevent orders to have 'impossible' due-dates assigned. A reference value for $h$ may be the minimum running time of an order. The BAT method can easily be implemented in an algorithm. We have a set $D$ of the size $q$ of assignable due-dates. Every route is represented with one due-date $d_r$ in the set. At the beginning of the simulation run the first departure times of every route are included in the set. As soon as the critical value of a batch is reached its due-date gets replaced in the set with the next departure time of the route (the next batch). Formally, with $t_a$ being the actual time in the simulation run we get:

$$D = \left\{ d_r \mid t_a + h \leq d_r < \left( \frac{t_a}{m_r} + 1 \right) \cdot m_r \right\}$$

An order is assigned with the probability $p_r$ to a route and thus its correlated batch and due-date. For practical use we suggest BAT for the simulation of systems which serve distant customers and serve transportation vehicles. The position of an intra-logistics system in a supply chain can be depicted by varying the number of routes $r$. For a system which is positioned in a high echelon of the supply chain it is more likely to have a larger number of routes than for a system in a low echelon such as a supplier of common resources such as coal or steel. For the simulation of systems which have to
supply in-house production the usage of RAN is suggested. The conditions in the systems environment (time pressure, predictability) can be illustrated by varying the interval out of which the flow allowances are selected.

C. Performance Availability Measurement

The basics of performance availability measurement according to VDI 4468 have already been explained in chapter II. Analogically, we divide our description into two parts, the measurement based on timeliness and the measurement based on waiting times. Basic parameters which have to be accounted for are depicted in Fig. 3. The required structures (e.g. resources and their states) should be integrated into the model at the inclusion of the measurement procedures at the latest.

![Fig. 3 Basic parameters for performance availability measurement](image)

The timeliness of an order can easily be determined within a simulation model. Required parameters to be collected are the accessing time of an order \( a_i \) as well as the completion time \( c_i \). The due-dates \( d_i \) can either be determined externally or internally while the simulation is running. Procedures for due-date creation have been examined in chapter B. The accessing time \( a_i \) is crucial for their determination externally as well as internally. Following the definition of [7] as depicted in (1) we define the number of tardy orders \( n_i \) at the moment of completion of order \( i \):

\[
    n_i = \begin{cases} 
    0 & \text{if } i = 1 \\ 
    n_{i-1} + 1 & \text{if } c_i > d_i \\ 
    n_{i-1} & \text{else} 
    \end{cases} \forall i = 1, ..., N 
\]  

(7)

with \( c_i < c_{i+1} \).

We have to alter the definition from (1) slightly and get

\[
    \eta_L = \frac{N - n_N}{N}. 
\]  

(8)

For any application the formulas can be treated as follows. Starting from a value of \( n_1 = 0 \), every time an order reaches its target later than its due-date the value of \( n_i \) is increased by one. If the completion time is smaller than the due-date no modification of \( n_i \) is performed. The \( N \)th (last) order serves the total number of tardy jobs and is the reference for the calculation of performance availability.

The measurement based on waiting times can be performed in a number of different ways. Reference [7] states that one has to specify business goals before making the decision of how performance availability will be measured. We believe that the specification of a single business goal cannot be recommended in a planning stage. Usually, a number of different business goals are considered. We therefore propose to measure all relevant waiting times and analyze the different resulting values with respect to specified objectives. Again, simulation serves the advantage of making the measurement of all waiting times comparatively cheap to a real-world system.

For further use, we specify two different types of waiting times to be measured. We define a main resource and supporting resources. The main resource is the one for which waiting times are measured. Supporting resources have the task of making the main resources operations possible. Their waiting times are not crucial but their operations ensure a proper flow of work. Supporting resources might be order tote slots, antecedent and subsequent conveyors or picking terminals. If, for example, the subsequent conveyor of a picking station is occupied and cannot receive a cleared storage tote the main resource is blocked and has to wait for the subsequent conveyor to be cleared. Hence, the states of the supporting resources have an influence on the waiting times of the main resource.

![Fig. 4 Case study system](image)
Following the guidelines of [7] the first way of measuring waiting times (henceforth called OT for overall time) is the recording of every time interval when the main resource is not occupied. For instance, there might be no task to fulfill or one of the supporting resources is not available. One will notice that this definition is simply the one of the utilization of a resource. Measurement procedures for this are already included in a large share of simulation environments and can be extracted easily. If not, the resources need to have the states ‘busy’ and ‘idle’ assigned. Every time the state changes to ‘idle’ the parameter beginning time $t_{b,k}$ is set to the actual time $t_k$. As soon as the state changes back to ‘busy’ the ending time $t_{e,k}$ is recorded as well. The overall waiting time after all tasks $k$ have been completed follows as

$$
T_w = \sum_{k=1}^{K} \left( t_{e,k} - t_{b,k} \right)
$$

and can directly be inserted to (2). It must be noted that the value of the waiting time between two tasks can very well be zero if the next task is instantly ready for processing. For practical applications a defined starting and stopping time for determination of $T_B$ must be defined. This can either be the starting and stopping time of the simulation run or the time of the first and last operation of the considered resource.

We define the second way of measuring waiting times (or SR for supporting resources) for interceptions which are provoked by the unavailability of supporting resources. For this purpose, the relevant supporting resources need to have assigned states which have the purpose of expressing the resources availability. The beginning time $t_{b,k}$ is set to the actual time $t_k$ if a task is ready for processing and the main resource is idle while the task cannot be fulfilled because one of the supporting resources is not available. This might be an empty order tote slot or occupied conveyors. The ending time will be recorded as soon as the state of the main resource switches to ‘busy’ (which indicates the task can now be fulfilled). Again, the overall waiting times can be calculated with (9) and inserted to (2).

Even though the ways of calculating the waiting times might interfere, two keynotes have to be created. As the times measured in SR are only a subset of the times measured in OT one would get faulty information by just adding the sums of OT and SR.

**IV. CASE STUDY**

With the following chapter we try to improve the understanding of the proposed guidelines for the conduction of a simulation study measuring performance availability. First, the system and its related data are described briefly. The resulting data is presented afterwards. Finally, some observations concerning performance availability are stated.

**A. Initial Situation**

For the measurement of performance availability we use the model of a small warehousing system with high work intensity. The simulation environment used is Demo3D by Emulate3D Ltd. The system consists of a two-aisle miniload, two picking stations, two packing stations and two putting stations which handle the incoming cartons. The subsystems are connected with a roller conveyor system. A screenshot of the system is provided in Fig. 4. The picking stations have two order tote slots each and provide space for one storage tote to be picked out at a time. For due-date creation BAT has been used. A total of six delivery routes need to be supplied, each of which has a departure interval of 60 minutes. The critical value $h$ was selected as 100 s which is slightly more than the minimum throughput time of an order. More relevant parameters can be gathered from Table I.

The simulation run covers 9 hours which is one whole workday. The model validation is conducted in two ways. Firstly, parameters such as miniload cycle times were recorded and compared to analytical results. Secondly, the processes were examined visually.

**B. Results**

The numerical results of the measurements are depicted in Table II. It is plain to see that they are twofold. On the one hand, the measurement of OT produces a low performance availability of 0.2 to 0.4 (depending on the resource). This is due to the characteristics of the input data which show a considerable amount of time intervals in which no tasks are left to complete. Thus, for this set of data we cannot recommend the usage of OT for contractual specifications. However, the figures indicate waste of resources at least for the number of putting stations. For the remaining resources a closer consideration for the behavior of the systems in peak periods needs to be conducted.

<table>
<thead>
<tr>
<th>Parameter Value</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. picking lines/hour</td>
<td>141</td>
</tr>
<tr>
<td>avg. orders/hour</td>
<td>46</td>
</tr>
<tr>
<td>avg. picking lines/order</td>
<td>3.05</td>
</tr>
<tr>
<td>avg. lines in prox 156</td>
<td>9.23</td>
</tr>
<tr>
<td>totes in stock</td>
<td>892</td>
</tr>
<tr>
<td>SKU-types in stock</td>
<td>453</td>
</tr>
<tr>
<td>incoming cartons</td>
<td>140</td>
</tr>
<tr>
<td>delivery batches</td>
<td>10</td>
</tr>
</tbody>
</table>

On the other hand, the measurement by means of SR as well as the measurement based on running times provides high performance availability.
values of performance availability. Certainly, comparing OT and SR, SR obviously serves lower values of waiting times as they are a subset of OT. Nonetheless we find it questionable to judge the performance of a warehousing system based on the waiting time of resources. Basically, a warehouse has the objective to provide desired articles at an appointed time at a specified location. The utilization of a resource is an index to consider but appears to be more of a question of cost effectiveness than performance (or availability).

The assessment based on timeliness is therefore the one we recommend. The creation of due-dates is a crucial step as it influences the outcome by a large degree. Hence, the creation has to be performed with care to achieve due-dates as close to the real-world system as possible. Also, the value of \( h \) has great importance for comparability when reporting simulation results (if BAT has been used).

C. Observations

In chapter II we propose the KPI lines in proximity \( p_{i,279} \). Having the results of the case study on hand we will try to justify its usage. We believe it provides a better reference to requirements of the system than indexes such as orders per hour. Firstly, the index does not compute an average for a large portion of time but considers every order individually. Secondly, not only the requirements for the order itself but the probability of a jammed system is incorporated. For the assessment of the KPI we compute the lines in proximity for every order \( i \) and every \( P \) from 1 s to 1000 s. To evaluate the quality of the respective values of \( P \) the correlation factor between \( L_{i,279} \) and the throughput time of an order is determined. The results are illustrated in Fig. 5. The correlation factor increases significantly to a value of 0.6 at around a proximity \( P \) of 80 s. It then remains on a high level (with a maximum of 0.64 at \( P = 279 \) s) until it declines at a \( P \) of about 350 s.

\[ \text{Fig. 5 Correlation factor for proximity } P = 1 \text{ s to } 1000 \text{ s} \]

Intuitively, the correlation factor is high around a \( P \) of 156 s, which is the average throughput time of an order. The high values in the interval of 200 to 350 s appear to be surprising. We believe the reason is that for high throughput times, the current system state has to contain congestions. Hence, orders which lie further in the future need to be considered. This is supported by Fig. 6 which indicates that the correlation is especially high for large numbers of \( p_{i,279} \).

\[ \text{Fig. 6 Correlation of lines in proximity } p_{i,279} \]

V. CONCLUSION

In this article we propose a paradigm for the creation of simulation models which measure performance availability of an intra-logistics system. Performance availability as a measurement for the degree of business objective fulfillment answers the demands of today’s intra-logistics systems assessment by a high degree. Thus, we propose to establish an international focus on the topic. We present axiomatic elements to consider when measuring performance availability in a simulation model. Certainly, for the establishment of certified simulation models our work requires extension, such as a qualified workflow or validation procedures for existing models.

REFERENCES