Scientific Research Publishing

American Journal of Operations Research, 2022, 12, 19-63
https://www.scirp.org/journal/ajor
ISSN Online: 2160-8849
ISSN Print: 2160-8830

Exploiting Virtual Elasticity of Production Systems to Respect OTD—Part 3: Basic Considerations for Modelling CPPS Characterized by Non-Ergodic Order Entry and Non-Deterministic Product-Mix for Fully Flexible Addressable Workstations

Bruno G. Rüttimann[1], Martin T. Stöckli[2]

[1]Eidgenössische Technische Hochschule Zürich IWF, Zurich, Switzerland
[2]Inspire AG, Zurich, Switzerland
Email: bruno.ruettimann@inspire.ethz.ch, stoeckli@inspire.ethz.ch

How to cite this paper: Rüttimann, B.G. and Stöckli, M.T. (2022) Exploiting Virtual Elasticity of Production Systems to Respect OTD—Part 3: Basic Considerations for Modelling CPPS Characterized by Non-Ergodic Order Entry and Non-Deterministic Product-Mix for Fully Flexible Addressable Workstations. *American Journal of Operations Research*, **12**, 19-63.
https://doi.org/10.4236/ajor.2022.122003

Received: January 6, 2022
Accepted: February 22, 2022
Published: February 25, 2022

Abstract

The recently experienced hype concerning the so-called "4th Industrial Revolution" of production systems has prompted several papers of various subtopics regarding Cyber-Phdysical Production Systems (CPPS). However, important aspects such as the modelling of CPPS to understand the theory regarding the performance of highly non-ergodic and non-deterministic flexible manufacturing systems in terms of Exit Rate (ER), Manufacturing Lead Time (MLT), and On-Time Delivery (OTD) have not yet been examined systematically and even less modeled analytically. To develop the topic, in this paper, the prerequisites for modelling such systems are defined in order to be able to derive an explicit and dedicated production mathematics-based understanding of CPPS and its dynamics: switching from explorative simulation to rational modelling of the manufacturing "physics" led to an own and specific manufacturing theory. The findings have led to enouncing, among others, the Theorem of Non-Ergodicity as well as the Batch Cycle Time Deviation Function giving important insights to model digital twin-based CPPS for complying with the mandatory OTD.

Keywords

On-Time-Delivery, Production System, Industry 4.0, CPPS, IoT, Stochastic Arrival Rate, Non-Ergodic Process, Virtual Elasticity, Production Capacity, Nominal Mean Exit Rate, Theorem of Non-Ergodicity, Non-Deterministic

Product-Mix, TFL, AGV, Scheduling Algorithm, Digital Twin

1. Introduction and General Approach

It has to be stated right from the beginning that cyber-physicality "*per se*", *i.e.*, increased digital content added to physical objects, will not significantly influence the performance of the underlying manufacturing system [1]. The performance of a manufacturing system in terms of Exit Rate (ER), Manufacturing Lead Time (MLT), and On-Time-Delivery (OTD) is largely determined by the manufacturing mode [2]. This is the "modus operandi" of how the manufacturing system is operated, which is defined by the selected implementation principles [2]. Hence the importance to know how to design a performant production system, because the "physics" of manufacturing performance is largely given by how the production system is engineered and less by how it is automated. However, also due to the increased performance of simulation techniques to explore the dynamics of various manufacturing concepts, the need to develop an understanding of the real "physics"—perceivable as a manufacturing theory—has not yet emerged. Universities obviously stress the simulation approach rather than trying to understand the production intrinsic laws of manufacturing dynamics [3]. This paper gives a contribution to the development of such a law-based theory proprietary to manufacturing, which goes beyond applied queuing theory and linear optimization techniques of operations research.

This paper is based on the far-reaching results worked out in two recent papers from Rüttimann & Stöckli [4] [5] in which the post-optimality conditions of On-Time-Delivery (OTD) in a deterministic product-mix manufacturing environment were presented. Based on the manufacturing-fundamental Theorem of General Production Requirements (OTD Theorem) enounced in [2], in paper [4], two new corollaries associated with the OTD theorem were defined:

- First Corollary to the Theorem of General Production Requirements
  (Corollary of Post-Optimality or Virtual Elasticity);
- Second Corollary to the Theorem of General Production Requirements
  (Corollary of Strong and Weak OTD Solutions).

as well as in [5], two further corollaries:

- Third Corollary to the Theorem of General Production Requirements
  (Corollary of Post-Optimal Backlog Waiting Time);
- Fourth Corollary to the Theorem of General Production Requirements
  (Corollary of Ergodic Backlog Rescheduling).

and a norm of nominal ER as well as the:

- Lemma to the Third Corollary of the Theorem of General Production Requirements
  (Lemma of Ergodic-Independent Validity) [5].

These four corollaries to the OTD theorem set the frame of the post-optimality conditions for On-Time-In-Full (OTIF) deliveries in a deterministic product-mix environment. Generally, the OTD theorem can be considered to be the most important theorem of production science at all. It has the same importance such as producing specification-conform high-quality products for customers. Furthermore, it is possible to model analytically the OTD performance for certain manufacturing systems [2] [4] [5]. The importance of this theorem and the four corollaries is derived from the fact to set the conditions for satisfying the imperative observance of punctual deliveries within an increased network-based system of fragmented supply chains. The Internet of Things (IOT) will even further favor this tendency of integrating fragmented supply chains. However, although everyone talks about OTD and OTIF, the necessary and sufficient conditions for OTD and OTIF have never been researched before 2017 [2]. The fundamental importance of these four corollaries is directly linked to the observance of conditions for OTD or, even more specifically, of OTIF compliance, which complies with the manufacturing cardinal *SPQR axioms of production systems* (speed, punctuality, quality, return) [2]. Now, following the insights outlined in [3], in this paper the findings of [4] [5] as well as the basic theory developed in [2] are extended and, if possible, applied to non-deterministic product-mix manufacturing systems of non-ergodic order entry process characteristics of CPPS allowing at the extreme case batch-size one. This extreme flexibility and elasticity are one of the topics of a CPPS such as interpreted by the German "Industrie 4.0 Arbeitsgruppe" i4.0 action group [6] [7], a mixed academic and industry smart factory task force.

Several papers, e.g., [8] [9] [10], still deal with basic frame setting and the implementation of digital twins (DT) with discrete event simulation (DES) of production systems. Further papers even deal with Industry 4.0 (e.g., [11]) still explaining the concept at an introductory level. This exemplifies the confusion created by the fuzziness of the CPPS-topic. However, the resulting theoretic performance of fully flexible CPPS is not known, if not simulated. The reason why it is not known is simple: the topic has not yet been researched and no theory has been formulated yet [3]. Allegorically said, the presently applied knowledge-gaining discrete event simulation DES-approach resembles rather a hands-on "trial and error" result exploring try-in approach than to a systematic "DOE-like" scientific approach to find a representative model explaining the behavior. This is a manifestation of the desolate state regarding present research of basic manufacturing theory [3]. Indeed, the behavior and the performance of complex manufacturing systems are usually addressed by applying queuing theory models, linear optimization techniques, and explorative DES (experimental manufacturing). Instead of putting the attention on simulation, academics should focus on the analytical cognitive modelling of the performance of CPPS leading to develop theoretic manufacturing knowledge. This shall lead to a novel teaching approach to train knowledgeable engineers [3]. Therefore, we will treat the specific OTD topic not heuristically or empirically as usual by simulation, but will prepare for applying the first-order logic, *i.e.*, by a theorem-based cognitive approach to developing the understanding. The applied, so-called, predicate logic bases in this case on production–specific laws enounced for the first time in [2] with a special focus on lean JIT manufacturing systems, but with basic laws also valid for other types of manufacturing systems. If this should be not enough, additional laws can be formulated. This allows exploring a Cartesian understanding of the OTD topic also in a non-deterministic environment such as represented by envisaged CPPS. The formalized, law-based understanding is not only didactically helpful to students' learning, but it is also interesting to consolidate empiric knowledge of both, young and not yet experienced manufacturing engineers as well as experienced production managers [3]. This paper has therefore also a didactic connotation.

At the base of modelling a manufacturing system stands the importance of throughput speed and on-time-deliveries. The importance is already reflected by the wording "general production requirements" in the fundamental *Theorem of General Production Requirements* (*OTD Theorem*), giving the necessary and sufficient conditions for OTD, which can be expressed mathematically for a manufacturing line according to Equation (1):

$$\text{for}: SD[OR] > 0 \quad \begin{cases} \forall i: \inf\{ER_i\} > E[OR] \\ \forall n: BWT_Z + MLT_Z \le EDT_n \end{cases} \tag{1}$$

where OR stands for Order Rate (called arrival rate in queuing theory) and ER_i for Exit Rate (departure rate) of each processing step, MLT for Manufacturing Lead Time of producing the whole batch, EDT_n the Expected Delivery Time of customer order n for OTIF deliveries, and BWT the Backlog Waiting Time of the queued orders for a given process. The index Z in the second equation of Equation (1) indicates the order entry point of the process from which the customer perceives the lead-time, called customer visible time (CVT). Therefore, the MLT starting from the raw materials might differ from the CVT if entry point Z is not at the beginning of the manufacturing process, but starts from a semi-finished component stored in a supermarket. Typically, the entry point Z could be a customization operation buffered with upstream standard components of Kanban-managed supermarkets. Equation (1) summarizes analytically the necessary and sufficient conditions for a manufacturing line to deliver a commercial order on time according to EDT.

But the OTD theorem is not enough. Typically, in non-perfectly engineered manufacturing systems, where only the capacity aspect, but not the manufacturing mode is addressed, specifically in "batch & queue" (B & Q) systems, the phenolmenon of WIP-formation is arising. The topic of WIP-generation has been analytically modeled in [2] leading to the *Law of WIP* formation represented in Equation (2):

$$\begin{cases} WIP = \sum_i WIP_i \\ \dfrac{\partial WIP_i}{\partial t} > 0 : CT_i > CT_{i-1} \\ \dfrac{\partial WIP}{\partial t} = E[IR] - E[ER] \approx \dfrac{1}{TT} - \dfrac{1}{CT_b} \end{cases} \tag{2}$$

and translated into the:

Theorem of WIP (or Delay Theorem or Time-Trap Theorem)

Given is a sequence of production steps with each process step having a different cycle time CT. Each process step with a longer CT than its preceding step is introducing a delay with the consequence of an increasing WIP formation in front of this process step.

Such a process step is called a time trap; therefore, a process may have more than one time trap.

And its corollaries for which we refer to [2]. For CPPS the understanding of this Theorem of WIP is of essential importance to see the consequences which non-lean engineered production line generate on *PLT* and *ER*. This is evidence that linear programming solutions (focussing mainly on capacity exploitation) are not sufficient.

In paper [4], we analyzed the problem statement regarding the post-optimality of OTD for a mono-product manufacturing system, in paper [5], we extended the OTD problem statement to a deterministic multi-product manufacturing system. In this paper, we address the preparation of the maximum variability characterized by non-ergodic order arrivals, variable batch-size, fully customizable products even with batch-size 1 and alternative routings, *i.e.*, high elastic and flexible production systems. The scope of these three papers is best represented by the schematic representations in **Figure 1**.

The rigid dedicated 1-to-1 model for mass-production of **Figure 1** reflects best Ford's original assembly line for the unique Ford T-model. The flexible n-to-1 cellular manufacturing was introduced by Toyota and allowed the production of different, but similar components within the same cell (or TFL). However, the products manufactured within the cell should comply with the Theorem of Cell Product Congruency (or Linear Dependency Theorem) or at least to its Corollary of Imperfect Dependency [2]. The possibility of unlimited flexibility (n-to-m) and elasticity is the target of the i4.0 action group [6] [7]. The scheme of n-to-m

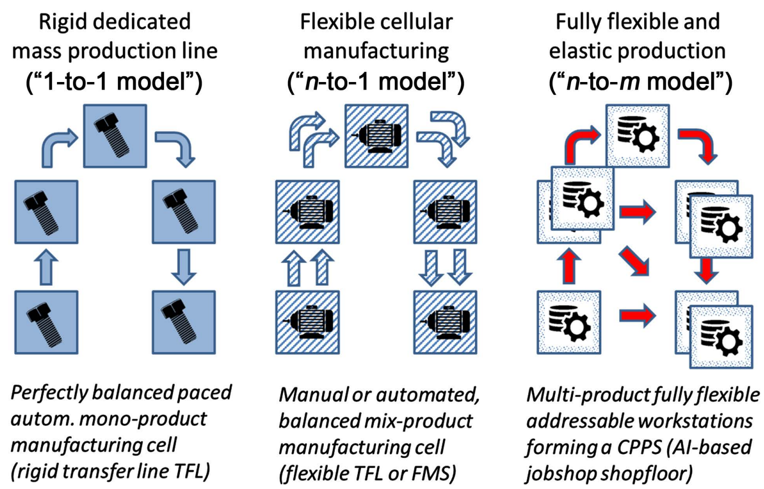


Figure 1. Schematic representation of manufacturing systems with increased flexibility and complexity with n products and m workstations.

shows the topology, or better the incidence structure, of a Directed Acyclic Graph (DAG) best suitable for modelling industry 4.0 CPPS. However, such an omnipotent CPPS is in practice a fully integrated system of automation based heavily on Artificial Intelligence (AI), generally speaking, applied to a merely common classic, widespread and often underperforming n-to-m jobshop model of shopfloor organization.

How can be explained CPPS in a few words? It is not the intention to deliver here a comprehensive definition trying to solve the created mess of unclear marketing buzzwords. For sure better than the coined and often used high-flying wordshell "4th Industrial Revolution" would be "Digitization based Large Scale Integration of Fully Flexible Production Systems"; but this clearly sounds less appealing. Important in this context is the word digitization. Toyota realized the integration of inbound and outbound with shopfloor logistics already 40 years ago—okay, only within a deterministic production environment and not with a stochastic environment. So what is the officially declared "main objective" of Industry 4.0 initiatives? Within the Smart Face project (Autonomik für Industrie 4.0) sponsored by the German "Bundesministerium für Wirtschaft und Energie" the following statement can be read "Development of a decentrally controlled small-scale production of electric vehicles based on Industry 4.0 technology & concepts" [12]. Autonomously Guided Vehicles (AGV) with intelligently identified semi-finished products interconnected with flexibly addressable workstations communicating via IOT are the basic underlying idea. Therefore, a theory derived from graph-based topology is the logic result of Industry 4.0-type conceived production systems applied on the shop floor to model CPPS [1]. According to the i4.0 action group [6] [7], one of the aims is to produce batch-size one not only intended as a standard customization possibility of a product, but as a one-off product to be able, stating literally, "to put a Porsche seat into a VW" [6]. This aims at allowing a production system to produce a stochastically variable product-mix, a "horror scenario" for every production manager. This characterizes the CPPS mainly as a Batch & Queue (B & Q) production based on a make-to-order production principle with push manufacturing principle to use the classification of [1]; the OR can then be associated to a non-ergodic order arrival process even extended to the shop floor. A fully stochastically variable product-mix entails multiple complications and is the opposite of Toyota's "panta rei" Mura-reducing philosophy trying to avoid any unevenness in production. The innovative JIT concept is therefore vanishing and such types of graph-based CPPS have to be seen as a step backward. Whether a fully stochastically variable product mix combined with unlimited order batch variability is desirable or even realistic, is not considered in this paper. Nevertheless, it has to be stated that this potential difficulty is nowhere mentioned by the i4.0 action group hereby standing in contrast to challenging the real difficulties to implement such an omnipotent system, or better, the consequence in terms of necessary equipment and capacity requirements as well as to the resulting performance of the production system. This is evidence for the lack of understanding not only the intrinsic aim of a JIT system, but also clear evidence of ignoring general manufacturing theory altogether [3]. Rather than expecting to see such an omnipotent system in the short-term, we will probably rather see very specific digitized solutions and restricted application of such CPPS, such as the Smart Face project itself represents, far away from the high flying "one-off" aim. This is merely a CPPS-disguised improved job shop solution under the "Industry 4.0" label. The manufacturing implications may even lead to the impossibility of implementing at the same time a high elastic (high variability and range of batch-size) and high flexible (high variability of product-mix) production system at reasonable costs [13]. This conflict of objectives has led Rüttimann and Stöckli to enouncing the Postulate of Incompatibility (or Flexibility-Elasticity Contradiction Postulate) [1] and has also prompted to write this paper. At least, the Smart Face project tries to implement the concepts on a small scale for which automotive application (single piece transfer principle) is ideal for testing instead of high volume and rather fixed TFL; the flexibility is intended to be implemented by addressable workstations served by AGVs. The description of a CPPS and a taxonomic comparison of different manufacturing systems are given in [1]. Detailed information about the intended CPPS can be retrieved at https://www.plattform-i40.de on the internet. This short description shows that with the so-called 4th Industrial Revolution and CPPS the intention and the focus is much more than only to display machine statuses virtually on tablets and smart goggles.

In this paper, we will not address the practical engineering aspects and conceptual implementation of such a complex system, but focus on the topic of theoretic modelling and performance characteristics of such systems.

The topic of general AI-based CPPS is presently highly researched. However, the research topics have an applicative connotation focused to introduce such cyber solutions rather than a theoretic shape of understanding such systems. Presently, its main applications are in the domain of predictive maintenance, which absolutely is not a new topic. Although certain elements of increased digitization have already entered modern shop floors, certain topics such as unlimited variability to manufacture one-offs like "to put a Porsche seat into a VW" [6] are presently rather on a wish-list. The authors, however, do not believe that unlimited variability makes much sense for a high volume flexible manufacturing line [14]. Also, the economic dimension of effort and result, or economic cost and benefit has to be covered. This opinion is also reflected in the already mentioned Postulate of Incompatibility (or Flexibility-Elasticity Postulate) [1]. It has to be clarified what is understood by flexibility and customization as well as about the extension of its meaning. The extreme flexibility paired with low volume or small batches is usually the business domain of SME with appropriate equipment and adapted organization to deal with such requirements. If MNEs enter this domain they will implement the "live and let die" paradigm of MNEs to the detriment of SMEs. The right degree of customization, and therefore required flexibility of the production, is essential also for elasticity consideration, *i.e.*, the impact on variable cost in the function of batch-size and total produced volume [4] [13]. Of course, it may also depend on the product: jeans vs cars. It is hard to believe that MNEs will enter niche markets on a larger scale, which poses the question not only about the strategic, but also about the economic understanding of the realistic requirements of present industrial logic by today's technology-minded and software-believing production engineers [13]. The degree of flexibility also depends on the product and the cost of customizing "one-offs". However, we will not question this and develop the mathematical logic to discuss this topic.

This paper's fundamental contribution lies in the importance to provide the basics for the topic of full product variability of so-called "4th industrial revolution" manufacturing systems with the first understanding of rational modelling regarding such systems. Therefore the understanding is derived from the application of solidly mathematically defined manufacturing laws based on [2] [4] [5] to model CPPS regarding the performance in terms of *ER*, *MLT* and finally OTD. Not gaining insight into the system's dynamics by applying situational case by case explorative simulation is the goal, but applying a general valid predicate logic of new formalized manufacturing theory to deduce the behavior. This allows for discussing and arguing about this topic based on rational and formalized findings instead of approximate and vague experience-based heuristics and personal preferences. The topic will not be exhaustively treated, but at least structured in order to allow Ph.D. students to carry out further research.

This paper is structured as follows: Section 2 introduces the topic of how to represent a CPPS mathematically in order to be modeled using the "objects" such as commercial and manufacturing orders, workstations, production capacity, backlog, and WIP, as well as the complexity regarding the concept of non-deterministic product mix with changing work content and non-defined manufacturing sequence. Section 3 defines the prerequisites and difficulties of a CPPS to attain OTD for all orders. The issues related to a dynamic changing bottleneck and an uncontrolled WIP generation are treated as well as the consequent necessary conditions of the system to respect OTD are developed. Section 4 enters the topic of performance of CPPS in terms of MLT, ER, and OTD. It addresses the characterization of a CPPS, the update frequency of the digital twin, and defines the laws to approach the required ergodic characteristics of production to have a performant manufacturing system.

2. Preliminary Assumptions and Special Characteristics of CPPS to Be Considered

The most important entities of a production system are the customer orders Ω for products k and the manufacturing system composed of machines, its manufacturing mode, and scheduling. In paper [5], we defined the entity order for a deterministic product-mix with:

$$\Omega_k := \left(\mathrm{X}_k, B_k, EDT_k\right) \tag{3}$$

where Ω_k stands for the commercial order for product *k*, characterized by a 3-tuple (tripel), having a product-specific order-rate distribution $X_k(\tau)$ with τ being the order inter-arrival time and reflecting the order frequency, order batch of size B_k, and an expected customer delivery time EDT_k of the product order to the customer. The specific product view is important also for the optimization of stocking strategies with classic "ABC-XYZ" techniques to manage the inventory of prematerial and components for every product according to demand and frequency. The manufacturing system, being characterized by a deterministic defined product-mix, has been represented in paper [5] by a non-necessarily quadratic matrix ***P*** of general dimension $m \times k$ showing the cycle times (CT_{mk}) for the k products on the m machines (workstations).

$$P = \left[CT_{mk}\right] = \begin{bmatrix} CT_{11} & CT_{12} & \cdots & CT_{1k} \\ CT_{21} & CT_{22} & \cdots & CT_{2k} \\ \vdots & \vdots & CT_{ij} & \vdots \\ CT_{m1} & CT_{m2} & \cdots & CT_{mk} \end{bmatrix} \tag{4}$$

We can define a norm on the deterministic Work Content (WC) for each product, corresponding to the machining or touch time of assembly or generally speaking to the necessary processing time, which is represented by the sum of the CT of the column vector of Equation (4).

$$\left\|WC_k\right\| = \sum\nolimits_m CT_{mk} \tag{5}$$

This is possible, because in a deterministic product-mix regime for each product order Ω_k exists an unequivocally defined and mapped association of the product k to its WC_k with manufacturing sequence (represented by the arrow):

$$\Omega_k \rightarrow \forall k : \exists \overrightarrow{WC}_k \tag{6}$$

In a fully flexible production system Equations (3) and (4) are not any more applicable, because the production system is now characterized by a highly non-ergodic Order Rate (OR) and a stochastically variable product-mix with variable Work Content (WC). Due to the fully variable WC of each product, the WC cannot be deterministically mapped in advance anymore according to Equation (6), but the WC defined in Equation (5) has to become part of the parametric order information. Therefore, the customer orders $\Omega_n = \{\Omega_1, \Omega_2, \cdots, \Omega_n\}$ is not characterized as Ω_k for the product k anymore, but generally to the generic entity "order" Ω_n becoming Ω a random variable which needs to include the product characteristics of WC. The concept of defined product becomes invalid and has to be substituted by the concept of parametric order. This is not only a semantic sophism, but a compulsory requirement if the final aim comprises fully customizable one-offs, *i.e.*, $B_n = 1$ or at least $\Omega_n \neq \Omega_{n+i}$ for $\forall i$. The stress is put on fully customizable and not until today only deterministically customizable from a catalogue of defined options. In such a case, the language of manufacturing should substitute the deterministic predefined concept of "sellable product" by the parametric concepts of "sellable competence-capacity" given by a defined range of manufacturing technologies leading finally to the “ad hoc” concept of “producible object”. Therefore, a commercial order translates into selling a variable sequence of manufacturing operations defined at the instant of the ordering; this becomes to be like providing a service. We will hereafter use the word “order” indifferently for commercial and manufacturing, however, when talking about batch, this is the translation of the order into manufacturing on the shop floor. By the way, the automotive industry is already in a batch-size 1 mode, however, the selectable optional features are deterministic and have to be chosen by a given set of predefined options (e.g., different engines, specific accessories, or defined colors) implemented with lean JIT-pull. And this is exactly what the i4.0 action group wants to change into full flexible one-off customization (“to put a Porsche seat into a VW” [6]). Now, as a consequence, Equation (3) modifies in Equation (7):

$$\Omega_n := \left(\overrightarrow{WC}_n, B_n, EDT_n\right) \tag{7}$$

and where the entirety of orders Ω_n are characterized by the random *OR* distribution written with capital Greek letter $\mathrm{X}_\Omega(\tau,t)$ with τ designating the inter-arrival time and t referring to the non-ergodic character. $\mathrm{X}_\Omega(\tau,t)$ is independent of the product *k*, because it is not existing deterministically anymore. This fact complicates also the buying and supply of raw-materials and necessary components. The necessary material might not be available when the order is placed or the procurement time is too long. The problem of synchronizing the availability of material arises again, topic which the lean TPS had solved. If we consider a manufacturing system with limited and deterministic customization possibilities on some or on all workstations forming a deterministic mix of product-categories, where *k* indicates not a product, but a product category, Equation (4) can be transformed into Equation (8):

$$\mathrm{P} = \left[E\left[CT_{mk}\right]\right] = \begin{bmatrix} E\left[CT_{11}\right] & CT_{12} & \cdots & CT_{1k} \\ CT_{21} & CT_{22} & \cdots & E\left[CT_{2k}\right] \\ \vdots & \vdots & E\left[CT_{ij}\right] & \vdots \\ CT_{m1} & E\left[CT_{m2}\right] & \cdots & CT_{mk} \end{bmatrix} \tag{8}$$

In the case of stochastic arrivals and *CT*, the resulting process lead-time (*PLT*)—*PLT* is the *MLT* for a single produced piece—can be calculated by applying Kingman’s and Kuehn’s approximation. However, queuing theory operates with the concept “lot” where the batch-size is not specified and usually intended to be one lot with *CT* referred to the entire lot, assuming a batch transfer principle. Indeed, Kingman’s and Kuehn’s approximation do not differentiate between a batch transfer or a single piece transfer principle and assume implicitly a batch transfer, limiting the application to B & Q systems. This shows exemplarily that the application of queuing theory remains on a very approximate level far away from the way how a real manufacturing lot transits modern shop floor layout with single piece transfer principle, if applicable. It makes a big difference if a batch B_k follows a $B_k\left(B_k\right)$, *i.e.* a batch transfer, a $B_k\left(n\right)$, *i.e.*, an n-piece transfer of the batch where n can be appropriately selected, or a $B_k(1)$, *i.e.*, a 1-piece transfer principle. If a "lot" is a "batch" with $B_k > 1$, then for a single workstation PLT and MLT will be equal, however, for a sequence of operations performed on different machines the difference can be huge [2]. Indeed, generally, in Western manufacturing theory was always applied a B & Q production philosophy with a batch transfer principle, reflected by queuing theory models, and not a "modern" fast SPF with single-piece transfer principle. This shows the lack of adequate mathematical modelling, a deficiency usually bridged by the help of Discrete Event Simulation (DES), but also the missing adaptation of generic mathematic queuing theory to meet specific needs of real modern manufacturing systems. We will exactly take care of this.

In this context, a further aspect to be observed is the "modus operandi" of the production system, *i.e.*, how the production system is operated and controlled. Different models exist with different shopfloor layout implications. A simple academically useful taxonomy of alternative basic production models was developed in [1] where the advantages with regard to flexibility and elasticity are compared. **Figure 2** shows two alternative shop floor layout of a CPPS: the first with a dominant "backbone" process forming the central transfer line (TFL) also apt for Single Piece Flow (SPF), and the second with an omnipotent "graph-based" full-flexible interconnected workstations served by AGV of rather job shop characteristics. The "backbone" TFL model is rather apt for customizable products belonging to a specified product category of rather defined manufacturing sequence. The "graph-based" AGV model is the only solution for a fully non-deterministic production allowing the manufacturing of a complete non-deterministic product-mix without restrictions, limited within the range of available technologies on the shop floor of course. Instead of AGV we could also name it AMR (autonomous mobile robots) which show even an increased flexibility compared to AGVs. However, we do not intend to enter the applied solution, but will only address the full flexible interconnection between workstations. The AGV/AMR are actually only apt for handling-optimized bins of batch and queue (B & Q)

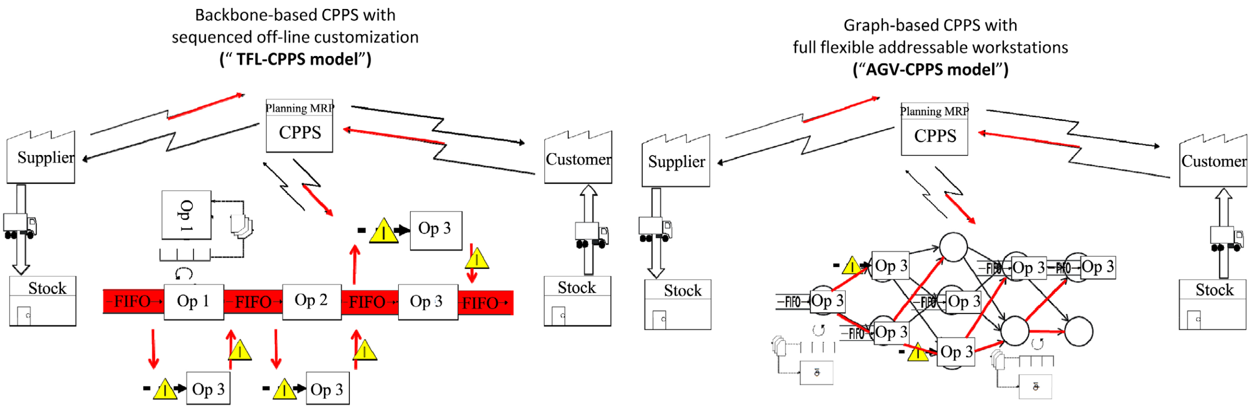


Figure 2. Alternative representation of CPPS shopfloor layout: way-optimized backbone-based "TFL-flowshop" vs full-flexible graph-based "AGV-jobshop" model.

manufacturing or SPF for very large objects such as automobiles. The predominantly applied manufacturing principle in both models will be "push". This is a direct consequence of the make-to-order production principle of a non-standard product. The flexible TFL-model layout does already exist in automotive applications, however with a yet limited AI application of the "cyber-content". The in-line flexibility of such existent solutions are limited to deterministic models however not taking customization of these models in consideration. With Figure 2 "backbone" CPPS, we mean the fully stochastic flexibility, where one-off customization might, however, realistically be executed rather off-line than in-line. For such manufacturing systems the OTD theorem according to Equation (1) fully applies, because no significant WIP will materialize if order release is appropriately managed.

Also the deterministic AGV-model exists already, however without decentralized neural controlling of the intelligent objects, represented by product, vehicle, and workstation. A separate paper would be necessary to compare the appropriate usage and advantages of each layout also in different contexts specifically such as manufacturing or logistics. The TFL-CPPS model should primarily schedule the backlog release and avoiding rescheduling of orders on the shop floor, which is even not necessary in a FIFO-lane TFL. The "backbone" orders should enjoy priority over the reentering of off-line customized orders, if not, loosing potentially a cycle for the whole line. The AGV-CPPS model allows also an extended rescheduling of order sequence directly on the shop floor to assure OTD. For this model, Equation (1) constitutes only the necessary, but not sufficient condition and will need to be modified due to the presence of multiple contemporary orders on the shop floor and the potential changing WIP of queued batches also caused by unbalanced execution of orders according to Equation (2). Both models need a fully parametric production matrix. The TFL-CPPS could also be modeled with Equation (8), which is ideally for fixed sequence, AGV-CPPS however not—the latter needs a complete parametric planning and dynamic rescheduling. From Figure 2 is evident, that also lean JIT-elements with Kanban managed supermarkets for components supply can be integrated in both CPPS models. However, defined supermarkets lose their "raison d'être" in stochastic mix and non-ergodic order regime manufacturing systems. Nevertheless, this shows clearly not "lean" or "4.0" is the question—they will go hand in hand [13]. Indeed, it is illusory to think that everything is instantly manufactured on demand (MOD), neither for repetitive-current identical nor for one-off components, perhaps with the exception of 3D additive printed manufacturing elements, but this production process is too slow for efficient industrial application. In the following, we will rather concentrate on the AGV-model. The TFL-model can be derived from the concepts already elaborated in paper [5], because the main order backlog (BL) of batches is located completely upstream of backbone TFL. Please note, also the TFL-CPPS model could be implemented physically via AGVs, however, the full flexibility can only be achieved with the graph-based CPPS. Indeed, as Figure 2 shows, the TFL-model is clearly designed around a specific type of products with imposed manufacturing sequences, whereas the AGV-model allows theoretically a fully variable configuration of the product in the sense of "Industry 4.0 Revolution".

In the following, we will not develop the intermediate step of Equation (8), because each variant of defined customization could be considered to be a product, and matrix of Equation (8) can be led back to Equation (4) with increased matrix dimension for the number of columns k where each column is a predefined variant. We intend to analyze generally what implies the full realistic or unrealistic aim of "ad hoc" variability of a non-deterministic one-off, such as intended by the German i4.0 action group, which cannot be put into a product category. Indeed, the aim of the i4.0 action group might be utterly unrealistic, but the problem solving of the problem statement is interesting from a theoretic modelling point of view. The customization, of course, has to be product and technology friendly to be able to be produced—we cannot mix manufacturing of cars and planes. Furthermore, we do not care about practical implementations to pass the WC to the workstations (such as the verification of the CAD model and the CAD/CAM transfer techniques). We are only interested in the theoretical modelling of a not yet existing theoretic CPPS also excluding temporary stand-still due to missing operators, or material, or breakdowns of equipment, as well as quality issues.

In the case of a non-deterministic product-mix, the matrix $\boldsymbol{P}$ (Equation (8)) cannot be defined in advance anymore. However, we anticipate that we could define a template matrix, let us call it production matrix, $\boldsymbol{P}_p$ with dummy elements representing free variable parameters of specific production times, *i.e.*, CT to manufacture one piece, at the different machines m (Equation (9)) for the n orders derived from the cartesian product of two sets $M \times \Omega(\Delta t)$ resulting of dimension $m \times n$ (see Equations (20) and (21)):

$$P_p = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & p_{ij} & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}_{\Delta t(\Lambda_n)} \tag{9}$$

The single dummy elements of the production template matrix $[p_{mn}]$ can be filled in, and updated within a dynamically recurring, but variable frequency $1/\Delta t$. This update frequency is the twinning frequency of digital twins. The matrix of Equation (9) will be valid within the timespan Δt which corresponds logically to the variable update frequency of entire order exit rate $ER = \Lambda_n$ or a campaign interval. This allows treating a non-deterministic problem statement with a deterministic structured modelling algorithm and formalism. We will resume later entering more in detail.

Equation (9) is the input template matrix representing the variable unitary resource absorption for the n orders usually applied in linear programming techniques. It does not tell us the manufacturing sequence of the n orders on the m machines. We have therefore to extend the concept of the order with the production sequence of addressed machines $\mathbf{\Pi}_{\Omega}\left[m_{ij}\right]=\mathbf{\Pi}_{\Omega}\left[m_i \to m_j\right]$ showing the transition map between the machines m_i and m_j for each single manufacturing order Ω. The sequenced work content $\overrightarrow{WC}$ is therefore expressed exhaustively by Equation (10):

$$\overrightarrow{WC}_n = f\left(WC_n, \Pi_n\right) = f\left(P_n \subset P, \Pi_n\right) \tag{10}$$

To capture the sequence of manufacturing (workflow), *i.e.* the sequential addressing of the various machines or operations to manufacture the products, we will base on graph-theory. Please note that we will not enter into the matter of graph-theory itself and refer to corresponding literature. The starting point of modelling the manufacturing sequence of an order is a non-directional graph showing the interconnection of workstations reflecting the topology of the manufacturing shop floor. We will assume for a CPPS a full connectivity of all stand alone machines (called vertices or nodes) allowing all combinations of the $m \times m$ graph-matrix with all element having the value 1. Instead of indicating the incidence or adjacency matrix (square matrix), which reflects only the connectivity of the knots (in our cases machines), we will take the transition matrix of Equation (11) representing a customizable template with an order variable topology (variable adjacency structure) of the graph (variable interconnection of the nodes which represents the machines) for each single order.

Like Equation (9) also $\mathbf{\Pi}_{\Omega}\left[m_{ij}\right]$ derives from a template matrix $\mathbf{\Pi}_{\pi}\left[\pi_{ij}\right]$ represented in Equation (11). The dimension of this transition matrix is according to the number of machines $m \times m$ of quadratic dimension and a dedicated matrix exists for each manufacturing order representing the ordered sequence the workstations are addressed. We could use for the state transitions a Markovian stochastic matrix indicating with increased probability the adjacent nodes of the manufacturing sequence. However, we prefer to use cardinal number to indicate the ordinal characteristic of manufacturing sequence. This transition template matrix—let us call it manufacturing matrix—where π_{ij} stands for the production transition between the workstations i and j when $i \neq j$ being usually $\pi_{ii} = 0$. The relative matrix positions π_{ij} will be occupied by ordinal numbers indicating the sequence of operations for the manufacturing order.

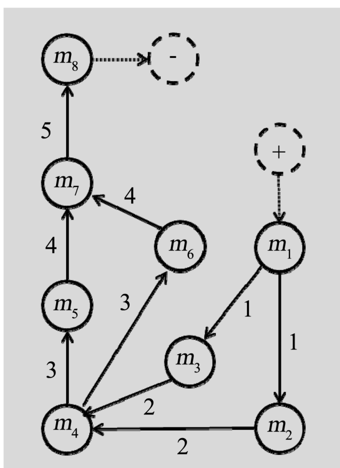
$$\Pi_{\pi} = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1m} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2m} \\ \vdots & \vdots & \pi_{ij} & \vdots \\ \pi_{m1} & \pi_{m2} & \cdots & \pi_{mm} \end{bmatrix}_{\Omega_n} \tag{11}$$

Equation (11) reflects the sequence on the topology of the graph-based representation of the shop floor. As shown by the matrix dimension $m \times m$, the maximum number of combinations is not given by the combinatorial calculation $\binom{m}{2} = \frac{m(m-1)}{2}$, but by $m^2 - m = m(m-1)$, because the transitions are bi-directional (bi directed graph). In the TFL-case Equation (11) will resemble to a sort of translated diagonal matrix $diag\mathbf{\Pi}_{\Omega}\left[m_{ij}\right]$ where $m_j = m_{i+1}$. In addition, the AGV-model allows also to select alternative manufacturing sequences, if possible, to optimize the load of workstations. The AGV-model is characterized by its topology. The topology of the graph gives the connectivity of process sequence with alternative "from-to" mapping of sequence. Here, we will not analyze the neural-based AI algorithm to allocate the orders to the machines, but only consider the topic of explicit mathematically modelling of the AGV-model interpretation of a CPPS. The AGV-model can be seen as the most flexible manufacturing system with omnipotent possibilities. In-between models, of course, are also imaginable and might be more realistic, at least at the beginning of the CPPS journey.

The manufacturing matrix of Equation (11) will be specific to each order Ω_n with the elements π_{ij} indicating the sequence of manufacturing for each single order; the elements are therefore characterized by ordinal numbers. The filling of the matrix bases on the linear sequence enumeration shown in Equation (12) which allows alternative paths.

$$\pi_\Omega = \left(\begin{cases} m_{a1} \\ m_{a2} \end{cases}, \begin{cases} m_{b1} \\ m_{b2} \\ m_{b3} \end{cases}, m_c, \begin{cases} m_{d1} \\ m_{d2} \end{cases}, m_e, m_f \right)_{\Omega_n} \tag{12}$$

In Equation (12) similar machines have been grouped. If the alternatives are limited we are in presence of a potential TFL. The concept of manufacturing sequence with alternative representation is shown in **Figure 3**. Please note that this has nothing to do yet with optimization of production order scheduling, which is a problem of the optimization type of sequencing the commercial orders to be manufactured. **Figure 3** shows only the sequence of logical operations for each single order.



$$\pi_\Omega = \left(m_1, \begin{cases} m_2 \\ m_3 \end{cases}, m_4, \begin{cases} m_5 \\ m_6 \end{cases}, m_7, m_8 \right)$$

$$\Pi_\Omega\left(m_{ij}\right) = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{\Omega_n} \end{array}$$

Figure 3. Alternative representations of the operation sequence for each production order. (a) Graph-based transition sequence (oriented topology description); (b) Set-based linear sequence enumeration (ordered set description); (c) Adjacency matrix-based analytical transition mapping (transition description with ordinal numbers referring to the sequence—not to the weights).

3. Additional Basic Theory Concepts Applied to Model CPPS

The most important concept within every constrained manufacturing system is the concept of the bottleneck, definition given by the Theorem of Throughput (or Bottleneck Theorem) [2], which is an important concept, because it defines the specific capacity of the production system. This theorem can be translated mathematically into:

$$\begin{cases} CT_b = \sup\{CT_m\} \\ ER = ER_b = \dfrac{1}{CT_b} \end{cases} \tag{13}$$

which states that the bottleneck workstation, defined as the workstation with the longest *CT* of a production line, limits directly the output, *i.e.* the exit rate (*ER*) of the whole production line. Equation (13) is generally valid and also shows the link between the Poissonian *ER*-view and the Markovian *CT*-view. In paper [5], a proto-norm for the ER for machine m and product k has been defined (Equation (14)), from which have been derived different aggregations of *ER* defining a space of normed nominal capacities.

$$\|ER_{mk}(t)\| = \lim_{n\to 1}\frac{dB_k(n)}{dt(n)} = \frac{1}{CT_{mk}} = \lambda_{mk}(t) \tag{14}$$

Equation (14) defines the *instant ER* (in queuing theory called departure rate) of the workstation m for the product *k* being manufactured. In the course of this paper, we will define a further nominal *ER*, *i.e.* an order-aggregated *ER*. The instant ER is relative to one piece produced and represents the specific capacity of the machine; the absolute capacity for a product is given by multiplying its *ER* with the shift model, *i.e.*, eight or sixteen hours. A flexible multi-product manufacturing cell is composed of several complementary and well sequenced-layout stand-alone workstations grouped into a cell forming a TFL "en miniature" able to produce a component. We underline the word stand-alone, *i.e.*, cell-process not aggregated e.g., into a single integrated rotary transfer machine, which we would rather associate with a perfectly paced extended workstation (integrated transfer manufacturing center).

According to **Figure 4** and Equation (13), the bottleneck for the two orders are apparently m_3 and m_4, respectively representing the longest *CT*. However, due to the fact that the other workstations are occupied by both orders, m_2 will likely become the bottleneck workstation of the production. The concept of bottleneck for several orders using the same equipment within a timespan Δt has to be extended to the concept of *WTT*.

For such a multi-product manufacturing cell the bottleneck workstation can be identified by calculating the workstation turnover time (*WTT*). According to the *Theorem of Generalized Throughput* (*or WTT-aggregated Bottleneck Theorem*) the workstation with the longest *WTT* is the bottleneck for this mix [2]. The $WTT_{m,k}$ is defined for the m^{th} workstation with the product-mix of *k* products as:

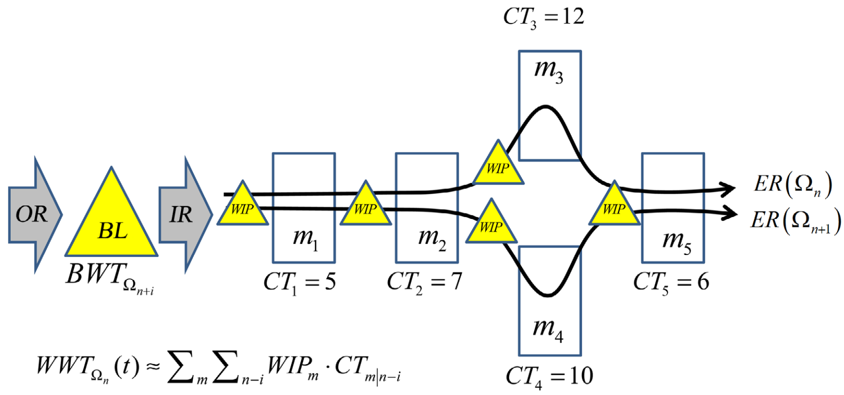


Figure 4. WTT (workstation turnover time) and WWT (WIP waiting time) are fundamental concepts in non-balanced multi-product manufacturing cells for several orders within a timespan Δt.

$$WTT_{m|k} = \sum_k \left(ST_{m|k} + CT_{m|k} \cdot B_k \right)$$

where $ST_{m,k}$ means set-up time of the workstation m for the k^{th} product and $CT_{m,k}$ is the cycle time of the m^{th} workstation for the k^{th} product and B_k is the batch-size of the k^{th} product. The WTT can be extended from specific product to the concept of generic orders $WTT_{m,\Omega}$ within a time interval. The introduction of WTT is required for a deterministic product mix to identify the bottleneck equipment. In the case of non-deterministic product mix, the bottleneck does not only change dynamically (see the corollary described hereafter), but the definition of WTT makes less sense, due to the difficulty to define the mix.

Several cells can be linked together via Kanban-managed supermarkets. A manufacturing cell constitutes a productive capacity usually suitable for deterministic LMHV (low mix high volume) production. For high variable mix including one-offs a technological competence is required as well as variable production paths between the equipment such as envisaged by CPPS-organised shop-floor, workstations served by AGV. The concepts of WTT and CTT not only lead to the Theorem of Generalized Throughput (or WTT-aggregated Bottleneck Theorem), but also to the First Corollary to the Theorem of Generalized Throughput (Corollary of Generalized Bottleneck Time-Variance) which is both valid within a mixed product manufacturing cell, and a common job shop-organized shop floor.

Within such a complex multi-cellular manufacturing system, the bottleneck cell is identified by calculating the cell turnover time (CTT). For the definition and calculation of CTT as well as further information consult [2]. We limit in this paper our discourse on AGV-CPPS to WTT. It has to be explicitly stated that the concepts of WTT and CTT make only sense in deterministic product-mix environment such as lean JIT-conceived manufacturing systems, or for a production matrix $\boldsymbol{P}_p(\Delta t)$ valid for a given time interval Δt, then WTT is calculated for the scheduled order mix in the campaign time interval. If ever, an approximate WTT or CTT could be defined for manufacturing systems according to Equation (8) with limited customization possibilities. Then we could imagine calculating an average expected $E[WTT]$ or $E[CTT]$ based on the expected distribution of type of order; however, we will not follow this approach. We will go directly to the core of the challenge represented by non-ergodic orders and stochastic product-mix AGV-type CPPS.

In the presence of multiple orders on the shop floor, such as represented in Figure 4, a WIP will materialize in front of each work station. The calculation of WWT (WIP waiting time) is not trivial, because it is dynamically changing when order Ω_n progresses along the downstream manufacturing process. The calculation of WWT is therefore only indicatively valid, theoretically at a given instant t. Figure 5 shows the generic situation for manufacturing order Ω_n in process at workstation m. At the next workstation $m+1$ the situation is of an order being processed on the machine and eventually already other manufacturing orders queuing, waiting to be processed on workstation $m+1$. The WIP_{m+1} of workstation $m+1$ will be fed not only by order Ω_n of workstation m, but potentially also from other workstations with an arrival rate AR. The departure rate (DR) from workstation $m+1$ will free the machine for another order waiting in the WIP_{m+1} to be processed.

The mathematics for the WIP dynamic in graph node m+1 is represented by Equation (15):

$$\begin{gathered}\frac{\partial WIP_{m+1}}{\partial t}=\sum_{\Omega} AR_{\Omega,m+1}-DR_{\Omega-1,m+1}=\sum_{n}\chi_n-\Lambda_{n-1}\\ \partial WIP=\left(\sum_{n}\chi_n-\Lambda_{n-1}\right)\cdot\partial t\rightarrow\approx\left(\sum_{n}B_n-B_{n-1}\right)_{\Delta t}\end{gathered}\tag{15}$$

where χ_n and Λ_{n-1}, respectively are the arrival and the departure rate, respectively of entire orders, *i.e.*, of the whole batch with order individual batchsize B_n and where χ_n (non-capital X_n) means potentially rescheduled, *i.e.* non-FIFO ordered. The remaining WWT at the workstation $m+1$ after Δt when order Ω_n arrives at the workstation $m+1$ is developed in Equation (16) (which is valid for all workstations hence the notation $m+i$):

$$\begin{aligned}WWT_{m+i}(\Delta t)&\approx\int_0^{\Delta t_i}\frac{\partial WIP_{m+i}}{\partial\tau}\cdot E\left[CT_{m+i}\right]\cdot\mathrm{d}\tau_i\\&=E\left[CT_{m+i}\right]\int_0^{\Delta t_i}\left(\sum_n\chi_{n,m+i}-\Lambda_{m+i}\right)\mathrm{d}\tau_i\\&\approx\left(\sum_n B_n\cdot CT_{n,m+i}-B_{n-1}\cdot CT_{n-1,m+i}\right)_{\Delta t}\end{aligned}\tag{16}$$

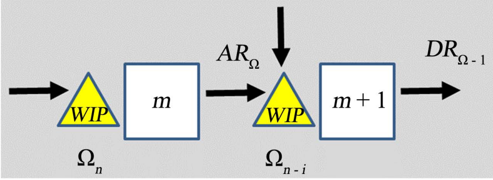


Figure 5. Arriving orders queuing as WIP in front of relative workstation. The WIP depends on the AR and the DR of the workstations.

where $\Delta t = B_n \cdot CT_{n,m}$ is the time the manufacturing order Ω_n is assigned at workstation *m*. A tentative algebraic description of the dynamic process of entire $WWT(t)$ valid at instant *t* for order Ω_n is represented by Equation (17):

$$WWT_n^{tot}(t) \approx \sum_m WWT_{n,m}^{FIFO}(t) \tag{17}$$

Equation (17) is a tentative representation valid only at the instant *t*, because when order Ω_n progresses along the downstream process the orders competing for the downstream workstations may change and consequently the *WIP* situation and WWT will change too.

The same reflection made for the WIP is valid for the backlog *BL* and its waiting time *BWT*. The arrival process is made by the non-ergodic order rate *OR*, which orders are stored temporarily in the *BL* before being released with the input rate *IR* to the shop floor best represented by **Figure 4**. The dynamics for the backlog is represented in Equation (18):

$$\frac{\partial BL}{\partial t} = E[OR] - IR \tag{18}$$

And the *BWT* can be calculated with Equation (19):

$$\begin{aligned} BWT(\Delta t) &= \sum_n \int_0^{\Delta t} \frac{\partial BL_n}{\partial \tau} \cdot CT_{n,b} \cdot \mathrm{d}\tau \\ &\approx \int_0^{\Delta t} \left(E[CT] \cdot E[OR(\tau)] - CT_b \cdot IR(\tau) \right) \cdot \mathrm{d}\tau \\ &\approx \sum_n B_n \cdot E[CT] - \Lambda_{n-1} \cdot \Delta t \end{aligned} \tag{19}$$

Equation (18) represents nothing else than the generalization of the first equation of Equation (1) with *TR* and *ER*. Like the first equation of Equation (1) also Equation (18) needs to be balanced in the average, *i.e.*, $E[OR] = IR$ in order to face a viable and ergodic order situation.

Now, the general OTD Equation (1) has unrestricted validity for a TFL or cell with one order in the system, but has to be adapted for modelling a graph-based, multi-product, multi-workstation AGV-CPPS with multiple orders in the system. A manufacturing system is mainly defined by the shop floor operations, represented by operators or machines *M*, and commercial orders Ω. Let *M* be a defined invariant set of addressable elementary stand-alone machines or workstations:

$$M = \{m_1, m_2, \cdots, m_m\} \tag{20}$$

and Ω a changing, dynamic set of variable elements corresponding to the commercial orders in a defined time interval Δt, however the set is of fixed dimension, *i.e.*, $\dim \Omega(\Delta t) = n$ where *n* = constant:

$$\Omega(\Delta t) = \{\Omega_1, \Omega_2, \cdots, \Omega_n\}_{\Delta t} \tag{21}$$

To each commercial order a manufacturing sequence with total work content is associated. The identification of the work content *WC* is not a topic of this paper as well as the technical data transfer and transformation of commercial order into a manufacturing-conform machine sequence. We characterize the $\widehat{WC}$ by an estimator-operator signalizing the fact that the WC will show non deterministic duration due to the variable and uncertain "prototype character" of potentially first time production introducing additional uncertainty to think at and therefore potential variation with immediate effects on the process dynamics [2]:

$$\widehat{WC}_n = \sum\nolimits_m \widehat{CT}_{mn} \tag{22}$$

Then, Equations (20), (21), with content of Equation (22), using the parametric production template matrix $\boldsymbol{P}_p$ of variable CT according to Equation (9) can form the real production matrix $\boldsymbol{P}_\Omega$ shown in Equation (23). This matrix will be kept fixed also in the dimension n (number of orders to cap the WIP) to comply with the Theorem of Lead Time Stability (or Steady State Theorem), applying the Lemma to the Corollary of Weak WIP Stationarity (Lemma of "Input Equals Exit" Principle), and limiting with this assumption the MLT variation. This lemma will be considered again for the twinning frequency and is usually referred to in literature with the CONWIP (constant WIP, *i.e.*, generic pull) implementation technique.

$$\hat{\mathrm{P}}_\Omega = E\left[\widehat{CT}_{mn}\right]_{\Delta t} = \begin{bmatrix} \hat{c}t_{11} & \hat{c}t_{12} & \cdots & \hat{c}t_{1n} \\ \hat{c}t_{21} & \hat{c}t_{22} & \cdots & \hat{c}t_{2n} \\ \vdots & \vdots & \hat{c}t_{ij} & \vdots \\ \hat{c}t_{m1} & \hat{c}t_{m2} & \cdots & \hat{c}t_{mn} \end{bmatrix}_{\Delta t(\Lambda_n)} \tag{23}$$

The generic element $\widehat{ct}_{ij}$ in Equation (23) of the matrix $\boldsymbol{P}_\Omega$ equals zero if no operation is scheduled on machine $M : m = i$ by order $\Omega_n : n = j$ or it has been already executed leading potentially to intra-twinning $\Delta t < \Delta t(\Lambda_n)$. The frequency of twinning Δt will corresponds to the entire order ER Λ_n to stabilize WIP with CONWIP or generic Pull technique; we address this topic later in more detail.

The OTD theorem, represented analytically by Equation (1), can be rewritten for the multiple linear equations of a CPPS (but also for a traditional manufacturing system) in compact matrix-form as represented in Equation (24).

$$\begin{gathered} \text{for} : SD\left[\left|\widehat{WC}_\Omega\right|_{\Delta t}\right] > 0 \\ \begin{cases} \hat{P}_\Omega \cdot b = r \le c \\ w + l \le d \end{cases} \end{gathered} \tag{24}$$

It is important to notice that the OTD theorem according to Equation (1) states that the necessary and sufficient conditions to be stationary (*i.e.* to face a variable OR) as well as OTD compliant (*i.e.* to observe punctual execution and delivery) are given by a capacity and a lead-time requirement observance. However, for a multi-product multi-order manufacturing system with limited resources (capacity of machines), Equation (1) gives only the necessary conditions, *i.e.* the restrictions to be observed. Due to the fact that several orders are present in the system on the shop floor and are manufactured simultaneously, competing for the same resources, of course in different stage of advancement, the target function is to have all orders of the campaign delivered on time. This challenge Toyota has easily solved with JIT Heijunka level-scheduled cellular manufacturing applying the "divide et impera" principle to the manufacturing system splitting not only the batches in pitches, but splitting also production in self-directed cells to achieve JIT. The solution of the target function of OTD-observance for all orders within a restricted capacity system of concurrent resources is the optimal scheduling sequence of commercial orders $\Omega_n^*(\Delta t_i)$ to be all on time within the time intervals Δt_i, which we will analyze in the next section. It is not guaranteed to have a strong solution for all manufacturing orders and the problem is not of solution uniqueness, but of solution existence. The concept of bottleneck becomes essential. An analytic discussion about the interpretation of bottleneck in Linear Programming and Lean Manufacturing has been made in [14]. Indeed, the specific capacity requirement stated the first equation of Equation (1) has to be extended when dealing with different products within a time interval of Δt. This because the specific capacity, which is a measure for instant performance, has to be substituted by the absolute capacity, where the effective available production capacity is reduced by the presence of machine setup time (ST) for the single orders.

Furthermore, until now we assumed to have a lean shop floor with SPF with no or at the limit negligible additional WIP if *CT* are balanced according to the Lemma to the Theorem of WIP, the Lemma of WIP Evenness (*i.e.*, with no WIP queuing, but only "WIP" being processed already taken into consideration by the *WC*). This assumption allows having the whole backlog in front of the first operation or better held back on hold to be released to the shop floor by an input rate (*IR*) according to **Figure 4**. However, in a multi-order batch-push operated job shop organized shop floor, as we have just seen, usually a WIP is forming with orders queuing in front of different workstations (**Figure 4**) leading to a *WWT* (WIP waiting time) according to Equation (16) causing a delay in execution increasing *MLT*. Hence, the recommendation to limit the release of orders to the shop floor, as it is recommended by the Lemma of Weak WIP Stationarity (Lemma of "Input Equals Exit" Principle). If not included into the *BWT*, the *WWT* has to be added to the *BL* orders forming the *BWT*.

Equation system 24 gives only the necessary, but not sufficient, conditions for the n orders to be on time, because it does not take into consideration the sequence of operations in the case orders are competing for resource assignment within a graph-based model. This leads to enouncing the

Fifth Corollary to the Theorem of General Production Requirements (Corollary of Extended OTD Conditions, or Multiorder WWT inclusion):

For a non-JIT, but graph-based organization of machines with multiple pushed orders on the shop floor with batch transfer principle queuing in front of the workstations, the OTD theorem represents only the necessary, but not sufficient, conditions for OTIF deliveries. For the sufficiency condition also the scheduling originated forming of sequenced WWT (t) *on the shop floor has to be included,* *however without assuring a strong OTD solution*.

The (t) in the $WWT(t)$ indicates the variability of WWT (WIP Waiting Time) along the order advancement on the shop floor. The WIP will materialize according to the *Theorem of WIP* and Equation (2) and Equation (15). Regarding the proof of the corollary, the reason for this corollary is the potential concurrent need for the same resource inducing a WWT for some orders queuing and waiting in front of the dynamically changing time-traps (or changing bottleneck), resource for which the classic MLT calculations for SPF (without WWT queuing time) have to be extended. This dynamic changing WWT can compromise OTD stating that this fifth corollary extends only to near sufficiency for OTD, necessitating and leading to enouncing the

***Lemma to the Fifth Corollary to the Theorem of General Production Requirements* (*Lemma of Potential OTD Solution Inexistency*):**

The presence of multiple orders on the shop floor competing for shared resources forming a dynamic WIP in front of the workstations can lead only to a weak OTD solution, *if ever*.

This is a further restriction to the Second Corollary to the OTD Theorem (Corollary of Strong and Weak OTD Solution) [4]. A weak solution means that the existence of a solution might depend on specific conditions, which have to materialize at the same time, but no general solution is available. The just enounced Fifth Corollary translates mathematically in Equation (25):

$$BWT + WWT + MLT < EDT \tag{25}$$

And written in terms of virtual elasticity:

$$BWT + WWT < \Delta T \tag{26}$$

Equation (26) represents the generalization of the Virtual Elasticity [4] for a production system with simultaneous multiple orders on the shop floor. The Third Corollary to the Theorem of General Production Requirements (Corollary of Post-Optimal BWT) [5] enounced for product order Ω_k extends the validity also to generic orders Ω_n of non-deterministic mix. We can use for the sum of MLT and WWT the abbreviation MLT^G to use the notation introduced in [2] reflecting the Universal Performance Law of Generalized Lead Time for Non-balanced Lines (for further information consult [2]) leading to Equation (27):

$$MLT^G = MLT + WWT \tag{27}$$

The second equation of Equation (1) and Equation (25) transforms into Equation (28):

$$BWT + MLT^G < EDT \tag{28}$$

with MLT for a batch push transfer principle calculated as:

$$MLT = B_k \cdot WC$$

The difference of nature regarding BWT and WWT stays in the type of queuing. BWT is before release to the shop floor, WWT is on the shop floor. It has to be stated that to reduce MLT^G the attention has to be directed to scheduling

BWT and not *WWT*; the *WWT* should be minimized in order to transit the shopfloor as fast as possible to achieve OTD.

Equation (24) has therefore to be extended on a multiorder jobshop shopfloor to Equation (29) by adding *WWT* and *ST*.

$$\text{for}: SD\left[\left|\widehat{WC}_\Omega\right|_{\Delta t}\right] > 0 \quad \begin{cases} \hat{P}_\Omega \cdot b + S \cdot 1_n = r \le c \\ w + v + l + \left(1_m \cdot S\right)^{\mathrm{T}} \le d \end{cases} \tag{29}$$

The first equation of Equation (29) implies that the time interval Δt to which the capacity $\boldsymbol{c}$ refers, the machines are always busy. The first equation of Equation (29) forms a linear system of m equations representing the capacity requirement of the OTD theorem. The time needed for the setup *ST* to change from one batch to the following order has been added, which can be considerable for non-lean optimized shop floors. The setup time *ST* is workstation-specific and therefore can be represented by a vector $s = \left[st_1, st_2, \cdots, st_m\right]$. Being ST_m only specific to each machine *m*, we will assume equal $ST_{m,\Omega}$ for all orders Ω_n on the same machine, the ST_m between machines can vary though, *i.e.*, formalized:

$$\text{for } \forall m, \Omega \quad \begin{cases} ST_{m,\Omega} = ST_{m,\Omega+1} \\ ST_{m,\Omega} \ne ST_{m+1,\Omega} \end{cases}$$

However, for matrix calculation purposes, the vector $\boldsymbol{s}$ of setup times is represented as a matrix $\boldsymbol{S}$ of dimension $m \times n$ taking the *n* orders into account (Equation (30)):

$$S = \left[ST_{mn}\right]_{\Delta t} = \begin{bmatrix} st_{11} & st_{12} & \cdots & st_{1n} \\ st_{21} & st_{22} & \cdots & st_{2n} \\ \vdots & \vdots & st_{ij} & \vdots \\ st_{m1} & st_{m2} & \cdots & st_{mn} \end{bmatrix}_{\Delta t} \tag{30}$$

with the single elements of $\boldsymbol{S}$ being assigned as follows:

$$\text{for } \forall(i,j): \text{if } \left[ct_{ij}\right] > 0 \quad \begin{cases} \text{then}: st_{ij} = st_i \\ \text{else}: st_{ij} = 0 \end{cases}$$

Further variables of the first equation in Equation (24) and (29) are the column vector $\boldsymbol{b}$ of order batch-sizes, which is a matrix of dimension $n \times 1$ according to the number n of orders; the vector $\mathbf{1}_n$ is a vector of dimension $n \times 1$ with all elements having the number 1 allowing to sum-up the lost capacities due to *ST*, transforming the matrix of dimension $m \times n$ into a vector of dimension $m \times 1$; and $\mathbf{1}_m$ of dimension $1 \times m$ a vector to transform the matrix of dimension $m \times n$ into a vector of dimension $1 \times n$. We will, however, omit for reasons of simplicity the set-up assuming by using the SMED technique (Single Minute Exchange of Die) or automation that the set-up time is negligible. Furthermore, the column vector $\boldsymbol{r}$ of resulting machine resource absorption, which is a matrix of dimension $m \times 1$; and the column vector $\boldsymbol{c}$ of maximum capacity given by the applied shift model, which is also a matrix of dimension $m \times 1$. The difference between $\boldsymbol{r}$ and $\boldsymbol{c}$ gives the shadow price of resources [14]. The shadow price will increase if specific machines are underutilized and corresponds to the opportunity cost. This is an important factor in CPPS targeting to manufacture also "one-offs" in special workstations with special machines leading potentially to costly overcapacities and the general difficulty to match OTD. It can result that the required EDT cannot be OTD matched and the effective resulting MLT indicates to renegotiate an acceptable EDT, unfortunately as it is often the case today.

The lead-time requirement of the OTD theorem is given by the second equation in Equation (24) and (29) represented by the column vector $\boldsymbol{d}$ of customer's expected delivery time ETD which is a matrix of dimension $n \times 1$. The column vector $\boldsymbol{w}$ represents the backlog waiting time BWT of queued orders waiting to be released to shop floor and the column vector $\boldsymbol{l}$ represents the manufacturing lead-time MLT of orders. This assumes that the BWT of the scheduled orders is at the beginning of the process to maintain the shop floor lean of orders (minimize WIP instead of WIP queuing in front of workstations). However, in case a WIP is not avoidable this is represented by the column vector $\boldsymbol{v}$ of WIP waiting time WWT.

The single vectors are represented in the transposed space-saving notation as follows:

Batchsize B	$b^{\mathrm{T}} = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}_{\Delta t}$
Unit column vector	$1_n^{\mathrm{T}} = \begin{bmatrix} 1_1 & 1_2 & \cdots & 1_n \end{bmatrix}_{\Delta t}$
Unit row vector	$1_m = \begin{bmatrix} 1_1 & 1_2 & \cdots & 1_m \end{bmatrix}_{\Delta t}$
Used resource	$r^{\mathrm{T}} = \begin{bmatrix} r_1 & r_2 & \cdots & r_m \end{bmatrix}$
Capacity	$c^{\mathrm{T}} = \begin{bmatrix} c_1 & c_2 & \cdots & c_m \end{bmatrix}$
BWT	$w^{\mathrm{T}} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$
MLT	$l^{\mathrm{T}} = \begin{bmatrix} l_1 & l_2 & \cdots & l_n \end{bmatrix}$
WWT	$v^{\mathrm{T}} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$
EDT	$d^{\mathrm{T}} = \begin{bmatrix} d_1 & d_2 & \cdots & d_n \end{bmatrix}$

where $v_n(t) = \sum_m v_{m|n}$ to be determined according to Equation (17) at the beginning and afterwards $v_n(t+1)$ will diminish while advancing downstream as well as the corresponding $l_n(t+1) := l_n - \sum_{m+1} ct_{m,n}$ is the residual MLT to be performed. We omit the explicit modelling of residual WWT at this point. The $BWT_n(t_0)$ is of course only present until the orders enter the shopfloor and becomes then zero. The update of P_Ω has to be seen together with the twinning frequency. Equation (19) represents only the conceptual model with the update algorithm to be designed. The vectors have not been provided with the estimator-operator such as $\widehat{ct}_{mn}$, because of a consequence of variable CT and therefore subject to variation. Due to indirect causality and for writing simplification, we will omit the operator-sign for these variables.

The first corollary to the OTD theorem, the Corollary of Post-optimality or Virtual Elasticity [4] of Equation (26) can be developed as follows for a multi-order graph-based CPPS (Equation (31)):

$$\begin{cases} w+v+l \le d \\ d-l = \Delta T \end{cases} \tag{31}$$

where ΔT is a column vector of dimension n writing $\Delta T^{\mathrm{T}} = \begin{bmatrix} \Delta T_1 & \Delta T_2 & \cdots & \Delta T_n \end{bmatrix}$ and where the elements of the vector are written in capital letters in order not to be confounded with the Δt time interval of rescheduling frequency resulting in Equation (32):

$$w(t)+v(t) \le \Delta T(t) \tag{32}$$

Equation (32) reflects the *Corollary of Post-optimality* in matrix interpretation extended with the WWT, representing the virtual elasticity of a multi-order manufacturing system, *i.e.*, the maximum BWT, and the maximum WIP waiting time WWT of queued orders in front of the machines, for each order to be still on time at a given instant t. Indeed, when the order enters the shop floor Equation (32) reduces immediately to:

$$v(t) \le \Delta T(t)$$

and the observation of OTD depends only on WWT, the MLT has already taken into account considering the virtual elasticity. The variability of $\Delta T(t)$ is the result of potential variability regarding expected real MLT of manufacturing order advancement on the shop floor. **The important conclusion: by formulating the virtual elasticity explicitly, the problem is therefore reduced to manage the WWT of each manufacturing order.** The WWT is not known deterministically in advance due to dynamic rescheduling. Equation (32) is also valid for TFL-model (BL at the begin), but where $\boldsymbol{v}$ becomes ideally zero. For an AGV-model not observing the *Lemma of "Input Equals Exit" Principle* (BL rather distributed as WIP along the manufacturing sequence) the WWT might not be under control. In the following, we will introduce to the development of a law-based mathematical logic to understand the optimal OTD scheduling solution of orders Ω_n^* instead of applying exhaustive DES, which could be at the extreme limit of $\mathcal{O}(n! \cdot m!)$ complexity.

The target function, such as in linear programming exercises, does not exist explicitly in this type of problem statement, the solution, or alternative solutions, derived from Equation (32) is the mix of sequenced manufacturing orders $\Omega_n^*(\Delta t)$ in the time interval Δt in order that all commercial orders comply to the second equation of Equation (29) (condensed into Equation (32)). Only in the case of multiple solutions and optimization target function would come into play. If a target function is defined, this could be minimizing set-ups of machines. This means without target function we don't have a mathematical optimization problem (maximizing or minimizing), but a problem statement of mathematical solution existence, at least if we do not assign values to orders with missed OTD or with an additional target function such as minimizing setups.

4. Understanding and Modelling Performance of CPPS

To develop further basic manufacturing theory applicable to CPPS we have to look at different aspects, which has to be analyzed in a structured approach. In the following sections we will only introduce to the topics for gaining insight into the complexity, always oriented to the goal of respecting the OTD target. We will highlight particularities in the case of distinctive characteristics of TFL- and AGV-model with regard to the application of the theory:

- The first topic of the problem tackles the engineering and applied implementation principles: Product characteristic, order frequency, and the consequent appropriate design of the manufacturing system. Here the question is: by which design implementation principles is a CPPS characterized;
- The second topic prepares the basics for algorithm-optimization of order sequence scheduling: Workstation load and digital twin-based simulation of job scheduling for OTD. Here the question is: which WIP cap and twinning frequency is optimal for short MLT to match OTD;
- The third topic analyzes the "physics" of the resulting performance of the applied implementation principles: Order characteristic and manufacturing mode defines the "physics" of the CPPS. Here the question is: which resulting performance in terms of MLT, ER, and OTD can we expect.

We will apply theoretic considerations developed in [2] [4] [5]. The manufacturing theory proposed in [2] consists of production laws and implementation principles, which characterize manufacturing systems. Some CPPS concepts will be compared to the lean-JIT TPS, simply, because it is the prototype of an existing manufacturing system for a deterministic product-mix with a self-governing rationale. The self-directed control of Kanban pull-governed lean TPS stands in diametral opposition to the present Western ERP still push-controlled B & Q manufacturing systems with inferior performance. Certain findings for CPPS are also valid for today's B & Q job shop manufacturing systems with only limited cyber-physical content. This because it is based on the push AGV-model. Indeed, before starting with implementing a CPPS one should have understood the intrinsic behavior of actual manufacturing systems. Therefore, this paper has also an enlightening didactic character.

4.1. Product Characteristic, Order Frequency, and the Consequent Appropriate Design of the Manufacturing System

Together with the order characteristics, the morphology and complexity of the product largely determines the manufacturing system concept. The physical characteristic is mainly determined by dimension and weight. The customization possibility adds a further dimension, which cannot be neglected. On the one extreme we have widely customizable single cars and on the other extreme we have production batches of thousands of standard small screws. In between we have e.g. electro-mechanical components and modules, which are in principle also capable for full customer specification as well as typically standard design windows, but always in customized dimensions, or even ad-hoc engineered products such as mechanic constructions. **Figure 6** shows exemplarily a rough product characterization and a possible design principle comparison, which are presently applied for typical products.

The type of customization is linked to the possibility of adaptation. It can range from simple impression of logo or special colors (rather to be called personalization) to special performance (genotypic customization) or dimensional specification without or even with morphological modification (phenotypic customization) as well as a complete new design (ex novo, "make to engineer"), which can be limited to a small series. The type or degree of customization can therefore be structured as follows:

- Imprint of logo or mere aesthetic feature (phenotypic personalization);
- Adaptation of characteristics and/or performance (genotypic specification);
- Ex novo designed (extended engineering).

The degree or extension of customization is also linked to the characteristic of flexibility of the shop floor layout, namely TFL-based CPPS versus fully flexible AGV-based CPPS such as intended by the German *i4.0 action group*. This leads to the necessary distinction between two extreme cases of batch-size 1 ("one-off") in industrialized production:

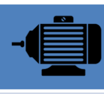
Product type					
Product characteristic	Standard of the shelf	Usually standard	Complex	Standard customized	Special
Morphology	rather small	handable	large	variable	wide range
Mix	Catalog deterministic	Catalog deterministic	Catalog deterministic	Catalog deterministic	Non deterministic
Customization	Standard, made to design possible	Standard, made to spec possible	Combination of many options	Always to customer specification	Design-specific construction
Frequency	recurring massproduction	from one-off to massproduction	variable massproduction	variable massproduction	variable
Production principle	Make-to-stock	Make-to-stock, make-to-order	Make-to-order	Make-to-order	Make-to-order
Nb. of components	single	subelements	many complex	standard elements	variable contingent
Nb. of steps manufacturing	reduced number	several with assembling	many multiple assemblings	several with assembling	variable contingent
Transfer principle	Batch or SP	Batch or SP	Single piece (SP)	Usually batch	Batch
Handling bin	bulk	on pallett	SP handling	on pallett	variable contingent
Manufacturing principle	usually Push	Push or Pull	Generally Push-Pull	Push	Push

Figure 6. Comparison of the presently applied main implementation principles for different product characteristics. High variability of special product type makes it difficult for automation.

1) Full customization of a deterministic mix (theoretically implementable);

2) Industrial manufacturing of fully non-deterministic mix (realistically impossible).

The case A reflects rather the TFL-model, which can be analysed stochastically and case B rather the AGV-model, which cannot be easily treated analytically anymore. An infinite customization of case A leads to case B. If not line-oriented, the AGV-model bases on functional departments with sufficient capacity, although organized as an "old fashioned job shop" it is interesting, because easily expandable and addressable. The classic AGV-model, however, contrasts to the lean principle of transport reduction, not only losing time for transportation, but also increasing inefficiencies due to potential WIP formation, if not managed, and necessary space occupation. In addition, the main manufacturing principle will be Push. New solutions such as "machine to operator" or "workstations aligning themselves according to flow" may become possible in future for certain situations.

The necessary effort of customization has to be seen in combination with the potential sales figures and the necessary investment to allow infinite flexibility. A viable Industry 4.0 "one-off" aspiration has to be seen within this framework [13], but also the realistic degree of modifiable engineered content. Indeed, the definition of "one-off" is linked to a combination of all three degrees of customization. The first successful IOT applications of e-Commerce are presently applied to personalization type such as e.g. jeans on demand, where not only the fashionable stickers, but also the design and size is fully variable. However, the concept of "putting a Porsche seat into a VW" [6] is still far away; we have to be realistic. This leads us directly to enouncing the

Postulate of Infinite Customization Impossibility (or Industrialized One-Off Illusion Postulate)

Although the AGV-based CPPS allows theoretically a full flexible shop floor production, the necessary investment to allow an automated producible full flexible non-deterministic order mix is subject to economic break-even considerations. The customization will rather be limited to product-specific changes applied to a specific product, mainly based on a product-specific TFL model.

This postulate foresees the impossibility to realize type B one-offs limiting customization only for type A deterministic mix one-offs. It has to be seen together with the Postulate of Incompatibility (or Flexibility-Elasticity Contradiction Postulate) [1]. These two postulates have to be considered as conjectures.

Furthermore, the characteristic of demand largely influences the applied production principle, *i.e.*, make-to-stock or make-to-order, which can also lead to a combined implementation. The Lemma to the Theorem of General Production Requirements (Lemma of "Make to X" Production Principle) [2] indicates whether a make-to-stock or a make-to-order production principle has to be applied. The reordering frequency of similar orders determines largely the availability of certain raw materials or components to be used in producing the required products.

The frequency of reordering, but also its ergodic or non-ergodic characteristic, is also an essential aspect to take into consideration, which, however, being an exogen variable, cannot be influenced. Influencible though, is the input rate IR from the BL to the shopfloor (**Figure 4**). Not anymore existing the concept of a fully deterministic product, but only existing the concept of the generic entity order we can only talk about the general order rate *OR* defined with generally non-ergodic *OR* characteristic $X_\Omega(\tau,t)$. For a non-ergodic process it may happen in a time interval Δt the most inconvenient case with Ω to have in sequence large WC_n and large B_n and short EDT_n. In case of a bulk arrival of such an order mix configuration, it might become impossible to deliver all orders on time. An alternation of small B_n with large B_n (and large WC_n with small WC_n) by applying the *Corollary of Ergodic BL Rescheduling* can render ergodic-similar the manufacturing order input rate. This leads to enouncing the *Theorem of* Non-Ergodicity:

Theorem of Non-Ergodicity (or Mura Theorem or Unevenness Theorem)

The variability of the orders can be given by a general non-ergodic arrival rate $X_\Omega(\tau,t)$ *combined by a very large standard variation of order-size* B_n. *Bulk arrivals of large orders can compromise the OTD requirement to supply all orders on time. Apart from the validity of the Corollary of Ergodic BL Rescheduling, a non-ergodic arrival rate only allows a weak OTD solution.*

This important theorem, together with the in [1] enounced Second Corollary to the Theorem of General Production Requirements (Corollary of Strong and Weak OTD Solutions), as well as the just before enounced Lemma to the Fifth Corollary to the Theorem of General Production Requirements (Lemma of Potential OTD Solution Inexistency), makes it impossible to have a strong OTD solution for B & Q push systems. This restriction relegates such a CPPS production system behind the Toyota invented JIT pull manufacturing system, which allows 100% OTD.

The Theorem of Non-Ergodicity addresses also the aspect of the order size, which translates into a manufacturing batch. Western philosophy of large batches to be supplied OTIF, which leads to the increase of inventory levels and *MLT*, stands in diametral opposition to lean pitch-supplied Japanese philosophy. Indeed, the *MLT* depends on the size of batch, number of operations, and the transfer principle, *i.e.*, batch or single piece [2]. It is preferable to have more frequent small batches, which we could call lots, rather than one big supply in order not to block capacity and to remain flexible. This reflects the Mura-concept with Heijunka-box levelling which allows replenishing the supermarket fast with all products or, generalizing, to satisfy several customers at the same time with short *MLT*. The "art" of scheduling is reduced to split the orders in suitable pitches (time slots of standard lot size), which guarantee all production orders to be supplied OTD. This will lead to enouncing a further corollary.

Already the Forth Corollary to the Theorem of General Production Requirements (Corollary of Ergodic BL Rescheduling) [5] can help to transform a non-ergodic OR into an ergodic IR, *i.e.*,

$$\mathrm{X}_{\Omega}(\tau,t) \rightarrow \chi_{\Omega}(\tau)$$

However, this is not guaranteeing the OTIF fulfillment for all orders, as the Theorem of Non-Ergodicity and the Corollary of Strong and Weak OTD Solution state. An alternative to the Corollary of Ergodic BL Rescheduling is to transpose the Heijunka pitch levelling of pull-implemented JIT production system to batch-push systems by fractioning the batches B_n of the order Ω_n into an adequate and standardized lot size B_n/l with limited variability, where l denotes the fraction to allow at least fractions of all orders to be supplied on time. This leads to enouncing the

Corollary to the Theorem of Non-Ergodicity (Corollary of Fractional Scheduling):

Aiming at a strong OTD solution by using the advantages of a Heijunka pitch-levelled production cell with fast MLT and limited variability maintaining flexible supply capability, the original batch size has to be split in standardized fractions, which allows overcoming non-ergodicity by approaching a deterministic planning situation.

This clearly contrasts the OTIF customer requirements, but allows at least partial OTD implementation with a supply-wide JIT production solution by minimizing WIP.

The number of necessary steps to produce an element for a component is a further point to be studied. It makes a difference whether the element can be produced, e.g., on a paced rotary transfer machine, leading to a defined *MLT*, or whether stand-alone, but inter-linked or freely addressable equipment is required, leading to all inconveniences related to scheduling of all processing steps within a push manufacturing principle based system. In addition, the transfer principle is mainly subject to the morphology of the component itself, but depends also from the engineers designing the production system and associated handling bins. Furthermore, the necessary physical variability of AGV to match the different morphology of the products (and the associated costs) poses a serious concern for unlimited mix variability in an automated manufacturing system, also for intelligent AMR.

A further aspect to be taken into consideration is the possibility to implement a pull manufacturing principle. **Figure 6** shows that presently push is often dominant. This is also linked to the fact that engineers not always know and understand the JIT pull concept well and even less how to implement it [3]. The question therefore is not "jobshop" or "flowshop" as often believed, *i.e.* batch or single piece transfer principle, but rather regarding the manufacturing principle, *i.e.* push or pull [2] [3]. Whenever possible, pull has to be preferred over push. It has to be precised that it is always possible to implement push, however a pull has to be well engineered. **Figure 6** also shows that pull is not necessarily limited to a make-to-stock production principle. Pull is rather linked to the concept of deterministic mix. This means that fully flexible CPPS have a trend to be push-based, which eliminates the advantage introduced by the TPS, though. This leads again to the question, which degree of customization is economically desirable.

As it emerges, the quest of omnipotent production systems seems to be limited to product-specific solutions. The omnipotent fully flexible manufacturing system such as postulated by the German i4.0 action group might remain a technology-driven vague dream of engineers, at least for the moment. This underlines the just enounced Postulate of Infinite Customization Impossibility.

Summarizing this section we have seen how to optimally execute an order, which translates in common language how to produce efficiently a product, depends on the knowledge of engineers to design a performant manufacturing system. This is all about the intelligent combination of the implementation principles [2] by knowing the advantages and the consequences of the resulting solution, and not primarily by blind brute force automation of an existing process. Mastering of the production specific theorems and corollaries, *i.e.* mastering the manufacturing theory, becomes therefore compulsory. For that, of course, the engineers have to know all the implementation principles and then to choose the combination, which delivers the best performance in terms of *ER*, *MLT*, and OTD—and then to automate. However, this is like a paradigm shift in present didactics and will even sound like a heresy for many manufacturing engineering course trainers and institutes who are focusing on automation and simulation. The new challenge is designing an intelligent manufacturing system and then to establish the economically appropriate automation degree.

4.2. Workstation Load and Digital Twin-Based Simulation of Job Scheduling for OTD

Section 4.1 defined the "hardware" and "operating system" how to conceive the manufacturing system. In this section we define the prerequisites of the "software" to control the "traffic", *i.e.*, the workload. Non-lean, *i.e.* non JIT-based, push-manufacturing systems need to be controlled by a MRP/ERP-system and optimized by advanced planning and scheduling (APS). This is a direct consequence of the complexity of a manufacturing system [2]. Today, the functionality of such production planning systems (PPS) are still limited to some built-in scheduling principles (e.g., FIFO, shortest processing time, SPT, earliest due date, EDD). To improve performance, they have to be enhanced by a software module called digital twin (DT) of the underlying manufacturing system allowing intelligent optimization of scheduling by simulation for targeted solution finding, usually targeted at OTD using APS software. The DT is a mirror image of the shop floor state, defined by statuses of machines and order advancements (**Figure 7**).

A shop floor DT representing the statuses of machines and orders allows a centralized optimization of n orders in the time interval Δt. This centralized approach is not according to the intention of the German *i4.0 action group*, which envisages a decentrally controlled neural system of intelligent objects [12]. However, we will omit the modelling of distributed neural intelligence and focus

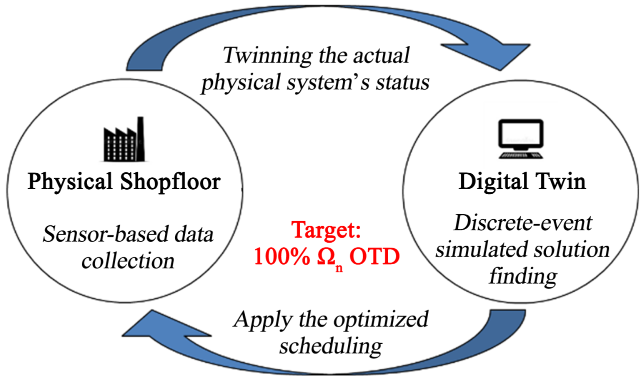


Figure 7. Optimization of scheduling for OTD by DT-based discrete event simulation.

on DT-based central optimization, a solution which is overall Pareto-optimal in the time interval Δt [3] and presently already applied. Such a system equals rather to "Industry 3.5". Indeed, we do not yet believe in the decentralized neural optimization approach for production planning and control because of the difficulty to achieve OTD for all orders due to the lack of the "whole picture". And even if the whole picture would be decentrally distributed, the question arises, which intelligent object gets the priority in the case of a resource conflict, or an arising problem clearly calling for central governance. Distributed intelligence leads to the necessity of a hypothetic Pareto-optimal decentralized algorithm, which perhaps is not implementable. We skip this issue.

Generally, DT need to be updated regularly, called twinning, to show a realistic mapping of the shop floor state. The twinning frequency $1/\Delta t$ can be chosen deliberately, but can also be optimized by being synchronized with events regarding the state-changes of the system. The DT update makes only sense in presence of a dynamic rescheduling possibility to exploit the full potential. Principally, the twinning of synchronization can be oriented in five ways:

1) $\Delta t(\text{fix})$: at a fixed interval (ideally matching a production campaign) or in the extreme case even continuously dt (what is called high frequency);

2) $\Delta t(\text{status change})$: each time when an order leaves a workstation (this allows reallocating the next logical order to the resource just become available in presence of exogen-imposed or endogen-required changes);

3) $\Delta t(\Lambda_n)$: each time when an entire order leaves the last operation and exits the shop floor (this could be used in addition to combine the triggering of the next order release from the order BL applying the "input equals exit principle");

4) $\Delta t(m_1 = \text{free})$: each time when the workstations of the first operation becomes again available (this has to be combined with the triggering of order release from BL);

5) A combination of 2) and 3) and 4).

This leads directly to define the Twinning Principle (at standard time, at any state transition, at order exit, at first free operation, a combination). Generally, it has to be decided how many simultaneous orders, *i.e.*, jobs, are present in the manufacturing systems, *i.e.*, on the shop floor. The decisional question is fundamen-

tal regarding the optimization to favor either:
- Maximization of apparent productivity or;
- Minimization of *MLT* [2].

Western manufacturing philosophy tends to flood the shop floor with orders maximizing productivity of the machines, forgetting—or rather not understanding—that the bottleneck determines the *ER*. Toyota-derived Japanese approach tends to have a lean shop floor favoring short lead times. In any case, for a B & Q operated system it has to result: the number of manufacturing orders has to be larger than the number of machines or workstations. The number of orders Ω_n, however, is not defined for a SPF operated concept, because the workstations are linked and would refer to the assignment of single pieces to the workstation of the same order. We can therefore enounce the

Theorem of General Workload (or Minimal Loading Theorem)

Given is a manufacturing system with m not further specified machines (workstations) and based on a batch transfer principle. The minimum number n of manufacturing orders in the system should be larger than the number of independently addressable workstations m.

Referred to the optimal dimension of the set in Equation (21), this translates mathematically to Equation (33):

$$\inf_n \{card\Omega_n\} \geq m \tag{33}$$

and the

First Lemma to the Theorem of General Workload (Lemma of Lean Shop Floor)

In the case of a production system based on the push manufacturing principle it is advisable not to flood the shop floor with manufacturing orders to minimize MLT and not to maximize apparent productivity.

Furthermore, for SPF we can enounce the

Corollary to the Theorem of General Workload (Corollary of SPF Load)

In the case of a manufacturing system with a single piece transfer principle the shop floor-released batch-size B_n of the order Ω_n should be larger than the number of workstation m. In the case $B_n < m$ additional orders can be released to the shop floor until the sum of the single pieces of the batches from different orders are larger than the number of workstations.

This can be translated with mathematical formalism as (Equation (34)):

$$\begin{cases} \inf_n B_n \geq m : B_n \geq m \\ \inf_n B_n < m : \sum_n B_n \geq m \end{cases} \tag{34}$$

The sequence of released orders has to consider OTD restrictions and, in addition, they can be released in function of minimizing costs e.g., optimizing set-up time. This is according to the intention to schedule the BL and not the WIP leading to enouncing the

Second Lemma to the Theorem of General Workload (Lemma of BL Scheduling Priority)

It is advisable avoiding a flooding of the shop floor with manufacturing orders that would require rescheduling the WIP to attain OTD for all orders, and to prioritize rescheduling of the order backlog over the WIP.

This lemma has to be seen in connection with the Lemma to the Theorem of General Production Requirements (Lemma of Flexible Scheduling Principle) enounced in [2]. The just enounced First and Second Lemma to the Theorem of General Workload are derived from the Lean theory, but can be applied equally to CPPS for maximizing shop floor performance.

Please note that the *C*orollary of SPF Load especially applies if the "after emptying set-up principle" is implemented [5]. In the minimal case of $n = m$ favoring short *MLT*, but eventually penalizing productivity, the resulting quadratic matrix would also offer the opportunity for matrix determinant calculation, whenever resulting to be convenient.

Having enough computational power, the continuous twinning principle of case a) seems to be ideal at a first sight, but does not take into consideration campaign-imposed technical restrictions (e.g., rolling aluminum sheets from wide to narrow widths and from soft to hard alloys). This indicates to favor a contingent approach of fixed interval, refreshing the twinning at the instant the last batch of the previous campaign begins the first operation. Here the interest consists of shortening the campaign horizon.

The twinning principle of case (2) allows a more frequent rescheduling, optimizing the scheduling sequence of present orders in the system. The twinning occurs just when the machines' status changes from "busy (or occupied)" to "inactive (or free)". The machine's status can be defined according to the Japanese Andon lights: green (busy = occupied), blue (setup = changeover), red (down = serious problem), orange (reduced availability = trouble), white (idled or inactive = free), ideally shortly before the job ends. The selection of the next order to be assigned to the available machine will be based on the combined evaluation of potential non-utilization time of the now available workstation regarding the next possible order transition to the available workstation to load it ("apparent importance" of productivity). This has to be done with the OTD observance of all orders necessitating this available resource in the imminent future in order not to compromise OTD ("essential importance" of punctuality), because when an order is assigned to the machine, pre-emptying is inefficient and costly. The twinning frequency f_{DT} depends on CT_m and B_n of the orders, which we can formalize in Equation (35). However, a higher twinning rate, such as suggested in different context in favor for high frequency twinning (case (1)), is not required in this production context and even not necessary.

$$f_{DT} = \frac{1}{\Delta t} = \frac{1}{\inf\left\{CT_m \cdot B_n\right\}} \tag{35}$$

Twinning case (3) is driven by bottleneck analysis and would suggest according to the *Bottleneck Theorem* [2] that Equation (35) transforms into Equation (36):

$$f_{DT} = \frac{1}{\Delta t} = \frac{1}{\sup\{CT_m \cdot B_n\}} \tag{36}$$

Equation (36) is correct for TFL-based production, but not generally valid for a AGV-based job shop systems if certain orders do not use the bottleneck resource. Indeed, the performance of an AGV system in terms of ER is also given by the load, *i.e.* the number of orders in the system. A high number of orders create WIP with the consequence of increasing *MLT*, but the WIP decouples the single machines, and having all workstations a WIP in front, they are all busy and can work at their own speed. Therefore, for the "load-approach" the BL is transformed as soon as possible into WIP. In addition, within a flexible shop floor with job shop organization, alternative manufacturing paths are possible even with different entry and exit points. This leads to further enlarge the concept of *ER*. In [5], a norm on the exit rate creating a space of normed nominal *ER* has been defined.

We will define a new *ER* of the entire production system **P** applied not to a single piece, but to an entire order Ω_n (or batch) designing it with capital lambda Λ_n formalized in Equation (37), which has to be added to the space of normed nominal ER writing:

$$\left\| E\left[ER_P\left(\Omega_n, t\right)\right]_{\Delta t} \right\| \triangleq \lambda_{\Omega_n}\left(\Delta t\right) = \Lambda_n\left(\Delta t\right) \tag{37}$$

This is important, because the bottleneck theorem represented by the second equation of Equation (13) can be bypassed if the bottleneck resource is operated in an extended shift model to work down the backlog. We have not to forget Equation (13) gives the specific capacity.

Twinning according to the *ER* of the entire order by applying the Lemma to the Corollary of Weak WIP Stationarity (Lemma of "Input Equals Exit" Principle) allows the capped WIP (defined number of manufacturing orders in the system) to remain constant [2] guaranteeing constant PLT. However, being the first operation maybe still busy (machine status "occupied"), the released order instead of waiting in the BL will queue in the WIP of the first operation.

At the instant an order leaves the shop floor, a column of the production matrix $\boldsymbol{P}_\Omega$ of Equation (21) becomes available and all parameters $p_{m\star}$ are 0, because no order is anymore assigned, a new order can be scheduled assigning to the $p_{m\star}$ the ct_{mn} of new order *n*. The status of the order on the shop floor can be: running (or in progress), waiting scheduled, waiting not assigned. How the communication between the intelligent objects such as orders (or AGVs) and machines regarding the conflicting situation of free resource assignment is solved, is not addressed in this paper.

The twinning case (4) goes against the "input equals exit principle" of order release triggering, although the WIP increase is negligible, it favors the occupation of first equipment and favors apparently the order advancement, apparently because of queuing later in the WIP. This has to be evaluated if the processing time of the batch B_n on the machine m_1 is larger than the WIP waiting time WWT of the orders Ω_{n-1} on the downstream machine m_2. In the case that:

$$B_n \cdot CT_{n,m_1} > \sum_{n-1} \Omega_{n-1}\left(B_{n-1}\right) \cdot CT_{n-1,m_2}$$

it may be opportune not to assign the order Ω_n to m_1 not compromising scheduling for OTD of orders waiting in the BL. This is also in line with the Lemma of Flexible Scheduling Principle stating "in order to keep until the latest instant the maximum flexibility in scheduling to allow preferential order treatment bypassing general FIFO principal of BL, it is advisable to release the next order into shop floor only when the first operation is imminent ready to accept it". We have not to forget that the concept of producing a standard product or producing a specific order is different such as it is best reflected also by the logic of pull-JIT SPF manufacturing from "pushed" B & Q is. Based on these considerations we are now able to enouncing the

Theorem of Twinning Frequency (or Rescheduling Theorem)

Given is a virtual image of the status of a physical manufacturing system. If no technical campaign restrictions exist, the minimal update frequency of this digital twin called twinning frequency f_{DT} *has to correspond to the optimal rescheduling frequency* f_{RS}. *The optimal rescheduling frequency has its logic maximum rate coinciding with the order departure rate from all workstations or any occurring state change due to exogen or endogen influences.*

Corollary to the Theorem of Twinning Frequency (Corollary of Campaign Matching Rate)

In the case of technological manufacturing restrictions forcing to campaign scheduling, the minimal update frequency of this digital twin has to match with the recurring production scheduling of campaign-imposed type of manufacturing.

This corollary can be formalized with Equation (38) as follows and reflects Equation (35) of case (2):

$$f_{DT} = f_{RS} := \left\{ \max\left\{\frac{1}{\Delta t}\right\} \middle| \Delta t = \inf\left\{CT_m \cdot B_n\right\}, \forall \Omega_{n|m}, t_i \right\} \tag{38}$$

The rescheduling frequency is also linked to the presence and intensity of noise: the exogen noise is usually customer-related; endogen noise is usually referred to machine break-downs or lack of material. Furthermore, the BL order release, *i.e.*, the input rate (*IR*), to the shopfloor has to be linked logically to the twinning cases (3) and (4), where the case (4) has to be favored in order not to apparently lose capacity at the first operation. This seems to contradict the widespread divulged concept of "input equals exit principle" leading to enouncing the

Lemma to the Theorem of Twinning Frequency (Lemma of BL Order Release)

To maximize apparent productivity and to minimize order lead-time, the order release to the shop floor can be triggered for orders not using the present bottleneck resource as soon as the workstation of the first operation for any or-

der becomes available, however, without strongly contradicting the "*input equals exit principle*".

Please note that this lemma has unlimited validity thanks to the diction "without strongly contradicting". Indeed, imagine a very fast first operation and a very slow, but dynamic changing bottleneck with $CT_1 \ll CT_b$. Then the risk exists to flood the shop floor increasing *MLT*. This lemma has its right of existence if the released order does not need the bottleneck resource. The "input equals exit principle" has still to be applied to minimize MLT to increase virtual capacity to assure OTD supplies for all orders. Not only due to technical campaign restrictions, the art consists of managing the BL and not the WIP rescheduling, especially if the setup times are not lean-optimized. The Lemma to the General Production Requirements (Lemma of Flexible Scheduling Principle) [2] exactly states not to flood the shop floor in order to remain flexible for OTD compliance. The art of engineering the manufacturing systems is intrinsic to allow fastest MLT, by applying the appropriate implementation principles. In the case of dynamic shop floor rescheduling, *i.e.*, if no fixed FIFO order scheduling is applied, and if the setup time *ST* tends to zero, then it apparently does not matter if the orders wait in the BL as backlog waiting time *BWT* or in the WIP as WIP waiting time *WWT* for the single order. However, the technical campaign restrictions still exist. This reasoning assumes a low variability of batch-sizes. The problem of AGV-CPPS is the potentially highly unbalanced workstations due to CT_m and/or B_n variability. Please note that the increased *WWT* may slow down the fastest MLT for further orders waiting in the BL reducing the virtual elasticity. This implies correctly applying the already mentioned Lemma of Flexible Scheduling Principle as well as the just above enounced Lemma of BL Order Release to optimize workload for achieving OTD. The First Corollary to the Theorem of General Production Requirements (Corollary of Post-optimality or Virtual Elasticity) [4] still remains valid.

4.3. Order Characteristic and Manufacturing Mode Define the "Physics" of the CPPS

Also CPPS will be governed by the theorems and corollaries already enounced in the text book [2] and the papers [4] [5] defining the "physics", *i.e.* the performance of a manufacturing system. However, to take graph-based CPPS specificity into account, new theorems and corollaries will become necessary to model the behavior of CPPS. Also for CPPS the speed-up of *MLT* is not only mandatory, but a priority. Such as the Lemma to the Theorem of Generalized Lead Time (Lemma of SPF Desirability) suggests, it is always opportune to try to install a SPF also in the case of unbalanced lines, because they exhibit a faster MLT than a batch transfer principle. Nevertheless, the aim remains to have a balanced line according to the Second Corollary to the Main Theorem of Production Time (Corollary of Balanced Line) [5], which is theoretically always possible to implement, but has to be engineered at the instance of planning the production sequence. Although in most cases a SPF will not be possible with a graph-based

AGV-model, the workstations occupied with batch jobs, *i.e.*, the occupancy, should be conceived to be balanced, in order to allow flowing instead of queuing and waiting, *i.e.*, implementing a "batch-flow". "Batch-flow" sounds to be an oxymoron, but has its semantic rationality, if no WIP is forming, imitating hereby a n-piece flow [2]. This makes it necessary to have a new and additional software module (workstation balancing software) performing the balancing of workstation occupancy taking into consideration the production order work content WC and the machine and workstation possibilities of the shop floor to design a batch-balanced work sequence. The Theorem of WIP defining the laws how WIP is generated keeps its fundamental importance and validity also for CPPS, but has to be adapted for batch transfer principle. The best would be to have an extended shop floor-balanced situation in place to limit WIP. In this case the balancing is not only related to balance the *CT* of a production line, but it becomes necessary to try to balance the occupancy of machines and workstations on the whole shop floor in presence of multiple orders with batch transfer principle. The focus is shifted from specific cycle time *CT* to absolute workstation occupancy time given by $B \cdot CT$ of the batch, which we can name batch cycle time *BCT*. This leads to enouncing for a batch-push manufacturing mode the important

Theorem of Balanced Workstation Occupancy (or Balanced Batch Cycle Time Theorem)

Given is a job shop-organized production based on batch transfer and push manufacturing principles. To avoid uncontrolled WIP generation between the operations, which would introduce a delay, the necessary condition to get a deterministically stabilized production requires that the batch cycle times BCT at all workstations have to be equal.

This theorem is linked to the *Second Corollary to the Theorem of WIP* (*Corollary of Strong WIP Stationarity or White Box Stationarity*) [2]. The just enounced theorem is represented algebraically in Equation (39) with a functional-implicit notation of the *BCT*.

$$\min\left\{\frac{\partial WIP_m}{\partial t}\right\} := B_n \cdot CT_{m,n} = const \,\middle|\, \forall m, n \tag{39}$$

The total, *WIP* is given by controlling the BL release *IR* to the shop floor, and therefore the *IR* has to comply with the *Lemma of* "*Input Equals Exit*" *Principle*. Equation (39) is in an extended sense Pareto efficient (**Figure 8**) and in case of inequality sign in Equation (39), a potential bottleneck is arising (Equation (40)) defining a dynamic bottleneck m_b, which dynamics is characteristic of highly non-ergodic, badly conceived production systems.

$$m_b := BCT_b = \sup\left\{B_n \cdot CT_{m,n}\right\} \tag{40}$$

The variability $SD\left[B_n \cdot CT_{m,n}\right]$ has to be limited dependent on the deviation of the average $E\left[BCT\right]$ from a target BCT_T as shown in Equation (41) to remain in a stochastically average balanced regime. The target BCT_T represents

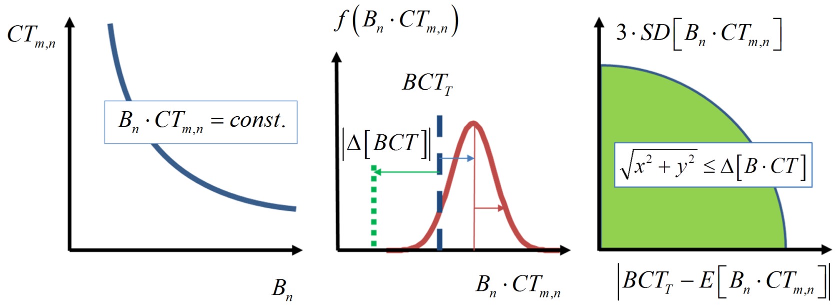


Figure 8. Relationship of batch size B and cycle time CT (left side) and the concept of the BCT Deviation Function (middle) as well as the domain definition of tolerance regarding maximum bias from average workstation occupancy $E[B \cdot CT]_{\Delta t}$ and maximum admissible standard deviation to approach a balanced load of workstations on the shop floor within a time interval $\Delta t \approx E[MLT]$ (right side).

a sort of specific batch capacity at each workstation to implement a sort of batch-flow, which will lead to enouncing a further theorem. To practically comply with the important just enounced *Theorem of Balanced Workstation Occupancy* a fixed tolerance ΔBCT has not to be exceeded for allowing a smooth operation not avoiding, but at least limiting WIP generation. Equation (41) can be considered to be the shop floor WIP-limiting condition for batch transfer principle balanced manufacturing systems:

$$\Delta[BCT] \ge \sqrt{\left|BCT_T - E[B_n \cdot CT_{m,n}]\right|^2 + \left(3 \cdot SD[B_n \cdot CT_{m,n}]\right)^2} \tag{41}$$

Equation (41) presents a similarity to the *Taguchi Loss Function* in quality theory and can be considered to be the equivalent in manufacturing theory in order to limit unbalanced operations and to have a predictable MLT for a standard production lot according to the formulas enounced in [2]. We can name Equation (41) the ***BCT Deviation Function***. Equation (41) is of central importance and translates the *Theorem of Balanced Workstation Occupancy* into applied reality. It can be summarized by the

Corollary to the Theorem of Balanced Workstation Occupancy (Corollary of Stochastically Minimized BCT Variability or BCT Deviation Function)

To limit WIP variation within a determined range, both, the centricity of batch cycle time BCT as well as the variation of BCT have to be confined within the domain of BCT tolerance to have a stochastically balanced workstation occupancy.

In its extreme case, the Corollary of Stochastically Minimized BCT Variability extends to a SPF with non-balanced CT in the process. However, the variation is simply reduced to the CT, which can be balanced out by applying the Second Corollary to the Main Theorem of Production Time (Corollary of Balanced Line). The just enounced corollary, together with the before enounced Corollary to the Theorem of Non-Ergodicity (Corollary of Fractional Scheduling) induces to an additional new implementation principle, the Pitch or Batch-size Principle (non-fractionized batch, standardized lot size, balanced fractional BCT). Only the non-fractionized batch principle allows OTIF supplies, at least for some orders. The standardized lot size principle might be imposed by standard handling bin constraints and as a consequence, it might be impossible to attain equal *BCT* on all workstations. Controlled by Heijunka-paradigmatic deterministic time-slot pitches, it is obvious that the optimal pitch principle to aim at is the balanced fractional *BCT* principle allowing potentially OTD of all lots. A lot is generally understood as a fractionized batch, but the naming is not applied consistently. Indeed, the *BCT* should rather be named Lot Cycle Time (*LCT*).

Referring to Equation (37), where a batch aggregated ER is defined and together with Equation (40) we can extend the CT-specific *Bottleneck Theorem* adapted specifically and valid for the batch transfer principle into the

Theorem of Generalized Batch Throughput (or Batch-Transfer Bottleneck Theorem)

Given is a non-stabilized job shop-organized manufacturing system with batch transfer principle. The instant bottleneck is given by the workstation with the longest batch cycle time BCT_b. *The average order-output of the production system is given by the average order exit rate* $\Lambda_n(\Delta t)$ *within a determined time interval* Δt, *where* Δt *cannot be smaller than the superior BCT time interval.*

The difference of this theorem to the WTT-Aggregated Bottleneck Theorem lies in the instantaneousness and general validity of this Theorem of Generalized Batch Throughput. The *WTT* is valid only within a certain time interval for a deterministic mix. For the Batch-Transfer Bottleneck Theorem, using for the related maximum *ER* of the production system for the orders Ω_n the capital Greek symbol Λ_n, we get Equation (42):

$$\begin{cases} BCT_b = \sup\limits_m \left\{ B_n \cdot CT_{m,n} \right\}_{\Delta t = B \cdot CT} \\ E\left[\Lambda_n(\Delta t)\right] = E\left[\sup\limits_{\Omega,m} \left\{ B \cdot CT \right\}_{\Delta t} \right]^{-1} \end{cases} \tag{42}$$

Equation (42) is the generalized batch interpretation of SPF-valid bottleneck Equation (13). The non-ergodic characteristic and difficulty to implement a pitch levelled production leads to enouncing the

Corollary to the Theorem of Generalized Batch Throughput (Corollary of Dynamic Bottleneck)

The variability of batch size B of each order and additional potential variability of cycle time CT at each workstation induces a variable batch cycle time BCT at each workstation generating a dynamic changing bottleneck of the manufacturing system.

The Corollary of Dynamic Bottleneck is the result of highly non-ergodic process on the shopfloor leading to potentially uncontrolled WIP generation. Non-ergodic processes on the shopfloor are a nightmare for production managers, a problem, which Toyota has solved reducing Mura (unevenness) with pitch-levelled Heijunka box scheduling. To stabilize production by limiting uncontrolled WIP generation in batch push manufacturing systems, the BCT Function Deviation has to be applied in order to approach the Theorem of Balanced Workstation Occupancy. The just enounced corollary justifies the mean operator of the order ER in Equation (42) for a batch-transferred order within a given time interval. The capacity of the system, however, is still given by the specific CT. The disadvantage of a dynamic changing bottleneck shows the fundamental importance of the just enounced Theorem of Balanced Workstation Occupancy and its corollary with the BCT Deviation Function to increase performance of CPPS.

According to the Lemma to the Generalized Lead-Time (Lemma of SPF Desirability), the goal is—if possible—to always implement a SPF [2]. Whenever possible it should be implemented also in CPPS by extending the concept of SPF to "batch-flow". A SPF strategy favors short MLT, but it might lead to a lower utilization of some off-line workstations. However, what might not be known, and that is of utmost importance, the ER at the end of the line is the same—if no equipment breakdown occurs, of course. Indeed, the Lemma to the Main Theorem of Production Time (Lemma of SPF Regime) recommends favoring a single-piece transfer also for unbalanced lines. The Corollary to the Theorem of Balanced Workstation Occupancy (Corollary of Stochastically Minimized BCT Variability) allows approaching a balanced workstation execution limiting WIP. This leads directly to enouncing the

Theorem of Batch-Flow Requirements (or Flow Imitating Theorem)

Necessary condition to speed-up manufacturing processes based on push and batch implementation principles is to strive for a balanced "batch-flow" by imitating an n-piece transfer principle, for which the Theorem of Balanced Workstation Occupancy has to apply.

A flow-like batch & balanced push system (B & BP) has a higher performance translated in shorter lead-time than classical batch & queue (B & Q) push systems. Indeed, the pure MLT without WIP can be calculated for an n-piece flow (nPF) according to Equation (43) [2]:

$$\lim_{n\to B}\{MLT_{nPF}\} = \lim_{n\to B}\{(B(n)-n+n\cdot m)\cdot CT\} = MLT_{B\&Q} = n\cdot m\cdot CT \tag{43}$$

The realistic lead-time has to consider also WWT, which can be approximated according to [2] with:

$$MLT_{nPF}^{G} = n\cdot\sum_{m=1}^{M} CT_m + \sup\{CT_m\}\cdot\left(B_k(n)-n+\sum_{m=1}^{b} WIP_{m,k-1}\right)$$

The aim is to split the batch in l lots of size n obtaining a fractionized batch B approaching an nPF of $B(n)=n\cdot l$, where n is the standard lot size and l the number of fractions. Equation (43) can then be approximated for the whole batch to Equation (44):

$$MLT_{nPF} = (m-1+l)\cdot n\cdot CT \tag{44}$$

where $n\cdot CT = LCT$. Equation (44) implies that the l lots forming the batch are produced subsequently, which jeopardizes the OTD of other orders. If $l=1$ then $MLT_{nPF} = m\cdot n\cdot CT$ which corresponds again to the batch MLT. More realistically, other lots from other orders will be alternated, which will, however, extend the lead-time to produce the whole batch.

To achieve OTD within a batch push manufacturing system, Equation (45) has to be satisfied, remembering that it should result $BWT + WWT \le \Delta T$, *i.e.*, shorter than the virtual elasticity, where $\Delta T = EDT - MLT$, *i.e.*, the validity of the First Corollary to the Theorem of General Production Requirements or Corollary of Post-optimality (Corollary of Virtual Elasticity) [4] extends to high variable non-ergodic orders with non-deterministic mix.

$$OTD := \begin{cases} \text{a)} \begin{cases} BCT_T \le E\left[\mathrm{X}^{-1}(\tau, t)\right] \\ \text{where}: \underset{n}{E}\left[\underset{m}{E}\left[B_n \cdot CT_{m,n}\right]\right] \le BCT_T \end{cases} \\ \text{b)} \begin{cases} MLT^G + BWT \le EDT \\ \text{where}: MLT^G = MLT + WWT \end{cases} \end{cases} \tag{45}$$

Although the Theorem of General Production Requirements (the OTD Theorem) is generally valid and therefore extends its validity also to B & Q systems, Equation (45) takes the specificity of batch transfer principle-based systems into account with $\mathrm{X}^{-1}(\tau, t)$ denoting the commercial order inter-arrival time. Equation (45) takes as well explicitly the WWT into account, stated in the Fifth Corollary to the Theorem of General Production Requirements (Corollary of Extended OTD Conditions, or Multi-order WWT inclusion). Equation (45) constitutes the generalization of the OTD Theorem and is valid for batch push CPPS. We can now enounce the

Theorem of Generalized General Production Requirements (or Batch-Generalized OTD Theorem)

Given is a production system based on batch-push manufacturing mode with a non-ergodic order arrival rate. The necessary, but not sufficient conditions to supply fractionized batches of the orders queuing in the backlog on time is to have enough capacity and a short lead-time. The necessary capacity is given by the condition that the target batch cycle time of the orders have to be smaller than the inverse of the mean order arrival rate. The admissible maximum lead-time including waiting time has to be shorter than the expected delivery time.

As the *Lemma to the Fifth Corollary to the Theorem of General Production Requirements* (*Lemma of Potential OTD Solution Inexistency*) and the *Theorem of Non-Ergodicity* state, Equation (45) gives only the necessary, but not sufficient conditions for OTD. Speeding-up MLT as well as reducing WWT by applying the *Theorem of Balanced Workstation Occupancy*, increases the virtual capacity of the system allowing a longer BWT for rescheduling the BL.

All these new theorems and corollaries as well as associated equations allow not only to calculate the performance of non-ergodic characterized CPPS, but also to engineer improved ergodicity-approaching shopfloor processes of CPPS. These laws are fundamental and form the foundation to conduct further scientific studies to improve the modelling of such manufacturing systems.

5. Conclusions and Remarks

This paper is far away from treating exhaustively the wide and complex topics of AI AGV-based CPPS modelling. It is a first tentative approach to structure the topic of modelling the governing dynamics and therefore, the optimized performance of omnipotent CPPS, *i.e.*, of new manufacturing systems envisaged by promotors of the so-called "Industry 4.0" initiative. Indeed, the topic of CPPS has to be considered to be at quite an embryonic development level characterized only by high-flying ideas without having captured the multi-dimensional complexity yet and still largely based on an experimental heuristics modus with a limited theoretic foundation—far away from being called the science. However, with the presented way here of modelling of CPPS and their performance—theorems, corollaries, and lemmas that are by the way also valid for non-cyber enhanced traditional B & Q manufacturing systems—the topic has become academically teachable. The Theorem of Balanced Workstation Occupancy with its corollary defining the BCT Deviation Function will become of central importance to conceive and optimize CPPS. By structuring the topic in theorems, students gain solid knowledge to understand the basic manufacturing "physics". The presented approach is an attempt to structure manufacturing theory in order to gain knowledge and to facilitate learning. The theory stands at the basis of didactics in order to pass knowledge about both physics and about experimental or empirical insights acquired by professionals especially to students for enabling them to conceive truly performant manufacturing systems.

Nevertheless, we have to make a last remark. It is often said that companies have to become lean before implementing CPPS. Although this statement is correct, many lean JIT concepts, such as the pull manufacturing principle or cellular manufacturing, are difficult to be implemented. However, non-value-add elimination such as waiting, piling, stocking, searching, remains a concern to be solved everywhere. Furthermore, big differences between cellular lean JIT systems and variable graph-based CPPS exist. Whereas lean JIT systems being self-controlled in theory do not need ERP systems for production planning and scheduling, CPPSs are centered on a digital twin-based optimization scheduling and controlling. On the other hand, Toyota's Kaizen continuous improvement approach for sure will evolve in CPPS to AI-based machine-learning, which will support agile sprints of self-directed continuous improvement teams.

This paper is unique in its kind in the domain of modelling non-ergodic orders with a non-deterministic mix in manufacturing systems of CPPS according to the German i4.0 action group and can be considered to be a seminal paper. It represents a starting point from which further research can be made, additional CPPS laws can be formulated and further papers can be written. The specific research opportunities regarding topics are e.g.:

- Specifications of ideal machines for AGV-optimized layout;
- Implementation and testing of the here presented CPPS modelling;
- Identification of optimized manufacturing path for shop floor orders in case

of conflict of resources;

- Development of the algorithm for balanced cycle time identification;
- Others.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rüttimann, B. and Stöckli, M. (2020) From Batch & Queue to Industry 4.0-Type Manufacturing Systems: A Taxonomy of Alternative Production Models. *Journal of Service Science and Management*, **13**, 299-316. https://doi.org/10.4236/jssm.2020.132019

[2] Rüttimann, B.G. (2017) Lean Compendium—Introduction to Modern Manufacturing Theory. Springer, Berlin. https://doi.org/10.1007/978-3-319-58601-4

[3] Rüttimann, B. and Stöckli, M. (2021) Considerations on Present Production Science Theory and Didactics: The Evident Lack of a Rational Manufacturing Theory. *Journal of Service Science and Management*, **14**, 482-501. https://doi.org/10.4236/jssm.2021.145031

[4] Rüttimann, B.G. and Stöckli, M.T. (2020) Exploiting Virtual Elasticity of Production Systems to Respecting OTD—Part1: Post-Optimality Conditions for Ergodic Order Arrivals in Fixed Capacity Regimes. *American Journal of Operations Research*, **10**, 321-342. https://doi.org/10.4236/ajor.2020.106018

[5] Rüttimann, B. and Stöckli, M. (2021) Exploiting Virtual Elasticity of Manufacturing Systems to Respect OTD—Part2: Post-Optimality Conditions for the Cases of Ergodic and Non-Ergodic Order Rate with Deterministic Product-Mix. *American Journal of Operations Research*, **11**, 141-165. https://doi.org/10.4236/ajor.2021.113009

[6] Kagermann, H. *et al.* (2013) Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0—Abschlussbericht des Arbeitskreises Industrie 4.0. https://acatech.de

[7] Bundesministerium für Bildung und Forschung (n.d.) Zukunftsbild Industrie 4.0. https://www.plattform-i40.de

[8] Jones, D., Snider, C., Nassehi, A., Yon, J. and Hicks, B. (2019) Characterising the Digital Twin: A Systematic Literature Review. *CIRP Journal of Manufacturing Science and Technology*, **29**, 36-52. https://doi.org/10.1016/j.cirpj.2020.02.002

[9] Ding, K., Chan, F.T.S., Zhang, X.D., Zhou, G.H. and Zhang, F.Q. (2019) Defining a Digital Twin-Based Cyber-Physical Production System for Autonomous Manufacturing in Smart Shop Floors. *International Journal of Production Research*, 57, 6315-6334. https://doi.org/10.1080/00207543.2019.1566661

[10] Park, K.T., Lee, J., Kim, H.-J. and Noh, S.D. (2019) Digital Twin-Based Cyber Physical Production System Architectural Framework for Personalized Production. *The International Journal of Advanced Manufacturing Technology*, **106**, 1787-1810. https://doi.org/10.1007/s00170-019-04653-7

[11] Asadollahi-Yazdi, E., Couzon, P., Nguyen, N., Ouazene, Y. and Yalaoui, F. (2020) Industry 4.0: Revolution or Evolution? *American Journal of Operations Research*, **10**, 241-268. https://doi.org/10.4236/ajor.2020.106014

[12] Schwede, C. and Gehrke, L. (2015) Smart Face—Industry 4.0 in Car Manufacturing.

Towards a New Paradigm for Production and Logistics. Fraunhofer IML/Volkswagen Group, Munich. https://www.iml.fraunhofer.de

[13] Rüttimann, B. and Stöckli, M. (2016) Lean and Industry 4.0—Twins, Partners, or Contenders? A Due Clarification Regarding the Supposed Clash of Two Production Systems. *Journal of Service Science and Management*, **9**, 485-500. https://doi.org/10.4236/jssm.2016.96051

[14] Rüttimann, B.G. (2015) Discourse about Linear Programming and Lean Manufacturing: Two Different Approaches with a Similar Converging Rational. *Journal of Service Science and Management*, **8**, 85-91. https://doi.org/10.4236/jssm.2015.81010