Forschungsberichte aus dem
wbk Institut für Produktionstechnik
Karlsruher Institut für Technologie (KIT)

Marvin Carl May

Intelligent production control for time-constrained complex job shops

Band 278

wbk
Institut für Produktionstechnik

Forschungsberichte aus dem
wbk Institut für Produktionstechnik
Karlsruher Institut für Technologie (KIT)

Hrsg.: Prof. Dr.-Ing. Jürgen Fleischer
Prof. Dr.-Ing. Gisela Lanza
Prof. Dr.-Ing. habil. Volker Schulze
Prof. Dr.-Ing. Frederik Zanger

Marvin Carl May

Intelligent production control for time-constrained complex job shops

Band 278

KIT
Karlsruher Institut für Technologie

wbk
Institut für Produktionstechnik

Intelligent production control for time-constrained complex job shops

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau des

Karlsruher Instituts für Technologie (KIT)

angenomme

Dissertation

von

Marvin Carl May, M.Sc. M.Sc.

Tag der mündlichen Prüfung:	21.12.2023
Hauptreferentin:	Prof. Dr.-Ing. Gisela Lanza (KIT)
Korreferent:	Prof. Tullio Tolio (POLIMI)

KIT
Karlsruher Institut für Technologie

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the internet at http://dnb.d-nb.de.

Zugl.: Karlsruhe, Karlsruher Institut für Technologie, Diss., 2023

Printed in Germany.

ISBN 978-3-8440-9464-0
ISSN 0724-4967

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

Vorwort der Herausgeber

Die schnelle und effiziente Umsetzung innovativer, nachhaltiger und wirtschaftlicher Technologien stellt den entscheidenden Wirtschaftsfaktor für produzierende Unternehmen dar. Universitäten können als "Wertschöpfungspartner" einen wesentlichen Beitrag zur Wettbewerbsfähigkeit der Industrie leisten, indem sie wissenschaftliche Grundlagen sowie neue Methoden und Technologien erarbeiten und aktiv den Umsetzungsprozess in die praktische Anwendung unterstützen.

Vor diesem Hintergrund wird im Rahmen dieser Schriftenreihe über aktuelle Forschungsergebnisse des Instituts für Produktionstechnik (wbk) am Karlsruher Institut für Technologie (KIT) berichtet. Unsere Forschungsarbeiten beschäftigen sich mit der Leistungssteigerung von additiven und subtraktiven Fertigungsverfahren, den Produktionsanlagen und der Prozessautomatisierung sowie mit der ganzheitlichen Betrachtung und Optimierung von Produktionssystemen und -netzwerken. Hierbei werden jeweils technologische wie auch organisatorische Aspekte betrachtet.

Prof. Dr.-Ing. Jürgen Fleischer
Prof. Dr.-Ing. Gisela Lanza
Prof. Dr.-Ing. habil. Volker Schulze
Prof. Dr.-Ing. Frederik Zanger

wbk
Institut für Produktionstechnik

Vorwort des Verfassers

Die vorliegende Dissertation entstand während meiner Beschäftigung als akademischer Mitarbeiter des Karlsruher Instituts für Technologie (KIT) am wbk Institut für Produktionstechnik.

Frau Prof. Dr.-Ing. Gisela Lanza gilt mein besonderer Dank für die Betreuung meiner wissenschaftlichen Arbeit als Hauptreferentin. Darüber hinaus bedanke ich mich für die Förderung und die zur persönlichen Entwicklung unerlässliche Rückmeldung sowie das Vertrauen, mir Gestaltungsfreiraum zu gewähren, der über das selbstverständliche Maß weit hinausgeht. Weiterhin danke ich Prof. Tullio Tolio für das Interesse an meiner Arbeit und die Übernahme des Korreferates, sowie Prof. Dr.-Ing. Kai Furmans für die Übernahme des Prüfungsvorsitz.

Dem InnovationsCampus Mobilität der Zukunft danke ich für die Förderung meines Forschungsaufenthaltes an der National University of Singapore. Mein ganz besonderer Dank gilt Frau Prof. Soh Khim Ong und Herrn Prof. Andrew Y.C. Nee und ihrem Team für die herausragende Gastfreundschaft und die vielen anregenden wissenschaftlichen Diskussionen.

Dem Karlsruher House of Young Scientists danke ich für die Förderung meines Forschungsaufenthaltes an der University of Texas at Austin in den USA. Ebenso gilt mein Dank Herrn Prof. Dragan Djurdjanovic und seinem Team für die herzliche Gastfreundschaft und die anregenden wissenschaftlichen Diskussionen.

Allen Studierenden, die zum Gelingen dieser Arbeit beigetragen haben, möchte ich danken. Besonders seien hier Lars Kiefer, Jan Oberst, Lukas Behnen und Sören Maucher genannt, deren Durchhaltevermögen einen entscheidenden Beitrag leisteten. Den Kolleginnen und Kollegen des wbk, vor allem im Bereich PRO, gilt zudem mein besonderer Dank. Der Gruppenzusammenhalt und die Möglichkeit zum Austausch haben diese Arbeit erst ermöglicht. Besonders seien die Kolleginnen und Kollegen des PSP-Teams erwähnt, deren tägliche Anteilnahme entscheidend zum Gelingen meiner Arbeit beitrugen. Sebastian Behrendt und Yannik Hermann danke ich herzlichst für ihre Akribie und Muße bei der Korrektur. Abschließend möchte ich mich bei meiner Familie für ihre unendliche Unterstützung bedanken.

Karlsruhe, 21.12.2023

Marvin Carl May

KIT
Karlsruher Institut für Technologie

Zusammenfassung

Im Zuge der zunehmenden Komplexität der Produktion wird der Wunsch nach einer intelligenten Steuerung der Abläufe in der Fertigung immer größer. Sogenannte Complex Job Shops bezeichnen dabei die komplexesten Produktionsumgebungen, die deshalb ein hohes Maß an Agilität in der Steuerung erfordern. Unter diesen Umgebungen sticht die besonders Halbleiterfertigung hervor, da sie alle Komplexitäten eines Complex Job-Shop vereint. Deshalb ist die operative Exzellenz der Schlüssel zum Erfolg in der Halbleiterindustrie. Diese Exzellenz hängt ganz entscheidend von einer intelligenten Produktionssteuerung ab. Ein Hauptproblem bei der Steuerung solcher Complex Job-Shops, in diesem Fall der Halbleiterfertigung, ist das Vorhandensein von Zeitbeschränkungen (sog. time-constraints), die die Transitionszeit von Produkten zwischen zwei, meist aufeinanderfolgenden, Prozessen begrenzen. Die Einhaltung dieser produktspezifischen Zeitvorgaben ist von größter Bedeutung, da Verstöße zum Verlust des betreffenden Produkts führen. Der Stand der Technik bei der Produktionssteuerung dieser Dispositionsentscheidungen, die auf die Einhaltung der Zeitvorgaben abzielen, basiert auf einer fehleranfälligen und für die Mitarbeiter belastenden manuellen Steuerung. In dieser Arbeit wird daher ein neuartiger, echtzeitdatenbasierter Ansatz zur intelligenten Steuerung der Produktionssteuerung für time-constrained Complex Job Shops vorgestellt. Unter Verwendung einer jederzeit aktuellen Replikation des realen Systems werden sowohl je ein uni-, multivariates Zeitreihenmodell als auch ein digitaler Zwilling genutzt, um Vorhersagen über die Verletzung dieser time-constraints zu erhalten. In einem zweiten Schritt wird auf der Grundlage der Erwartung von Zeitüberschreitungen die Produktionssteuerung abgeleitet und mit Echtzeitdaten anhand eines realen Halbleiterwerks implementiert. Der daraus resultierende Ansatz wird gemeinsam mit dem Stand der Technik validiert und zeigt signifikante Verbesserungen, da viele Verletzungen von time-constraints verhindert werden können. Zukünftig soll die intelligente Produktionssteuerung daher in weiteren Complex Job Shop-Umgebungen evaluiert und ausgerollt werden.

Abstract

In wake of an ever increasing complexity the desire to move towards intelligently controlling operations is amplified in manufacturing. Complex job shops mark the most complex production environments that require a high degree of agility to control. Among these complex manufacturing environments semiconductor manufacturing stands out as it combines all complexities to form a truly complex job shop. Hence, operational excellence is the key to success and relies on intelligent production control. A major concern in controlling such complex job shops, in this case semiconductor wafer fabrication, is the presence of time-constraints that limit the transition time of products between two, mostly successive, processes. Adhering to these product specific time-constraints is of utmost importance as violations result in scrapping the violating product. The state-of-the-art production control of these dispatching decisions that aim at adhering to time-constraints is based on error-prone manual control that is stressful for human operators. Thus, within this thesis a novel, real-time data based approach for intelligently controlling production control for time-constrained complex job shops is presented. Using an up-to-the-minute replica of the real system both uni-, multi-variate time series models and a digital twin are used to obtain violation predictions. As a second step, based on the time-constraint violation expectancy the production control is derived and implemented with a real-world semiconductor manufacturing plant real-time data. The resulting approach is, therefore, validated against the state-of-the-art showing significant improvements as many time-constraint violations could be prevented. In future, thus, intelligent production control should be evaluated and rolled out in more complex job shop settings.

Contents

Contents I

Abbreviations IV

Formula symbols VI

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Problem statement . . . 3
- 1.3 Research hypothesis . . . 4
- 1.4 Structure of this work . . . 5

2 Fundamentals 6
- 2.1 Semiconductor Manufacturing . . . 7
  - 2.1.1 Semiconductor fabrication technology . . . 8
  - 2.1.2 Frond-end wafer fabrication . . . 14
  - 2.1.3 Time-constraints in Semiconductor Manufacturing . . . 17
  - 2.1.4 Summary: Semiconductor manufacturing complexities and requirements 19
- 2.2 Production planning and control . . . 20
  - 2.2.1 Key Performance Indicators . . . 21
  - 2.2.2 Production Planning and Control tasks . . . 22
  - 2.2.3 Complex Job Shops and Production Planning and Control . . . 26
  - 2.2.4 Semiconductor Production Planning and Control . . . 26
  - 2.2.5 Summary: Solution approach requirements . . . 30
- 2.3 Quantitative optimization methods . . . 31
  - 2.3.1 Mathematical optimization . . . 32
  - 2.3.2 Heuristics and metaheuristics . . . 34
  - 2.3.3 Artificial Intelligence . . . 35
  - 2.3.4 Machine Learning . . . 38
  - 2.3.5 Predictions with Time Series Models . . . 48
  - 2.3.6 Summary: Quantitative methods to optimize production control . . . 51
- 2.4 Production system Digital Twin . . . 52
  - 2.4.1 Production system simulation . . . 52
  - 2.4.2 Foresighted Digital Twins . . . 53
  - 2.4.3 Knowledge Graph based Digital Twins . . . 54
  - 2.4.4 Summary: Simulations as Digital Twins for production control . . . 57

3 State-of-the-art literature review **59**
3.1 Literature review of focus areas 59
3.1.1 Digital twins for intelligent production control 60
3.1.2 Dealing with time-constraints in capacity planning and scheduling . . . 62
3.1.3 Adhering to time-constraints in dispatching in complex job shops 67
3.1.4 Implementing learning based production control in job shops for time-constraints 70
3.2 Research deficit 70

4 Intelligent Production Control for time-contrained complex job shops **76**
4.1 Problem scope and assumptions 77
4.2 Modeling the production system 78
4.2.1 Relevant system elements 79
4.2.2 Simulation and foresighted digital twin modeling 81
4.2.3 Transitional modeling approach 92
4.3 Intelligent production control architecture for time-constraint adherence 95
4.3.1 Control flow architecture 95
4.3.2 Intelligent production control for time-constraint gate keeping decisions 97
4.3.3 Implementation of time-constraint gate keeping decisions in operations 104
4.4 Transition time and adherence prediction 105
4.4.1 Time-constraint adherence prediction with foresighted digital twin 106
4.4.2 Time-constraint adherence prediction with transition model 108
4.4.3 Obtaining Prediction Intervals for time-series 111
4.5 Performance evaluation for prediction and prediction interval benchmarks . . . 124
4.5.1 Foresighted digital twin computational performance 124
4.5.2 State-of-the-art and predictor benchmark performance 126
4.5.3 Prediction interval model evaluation 129
4.6 Summary of the overall approach and framework 132

5 Evaluation and computational results **136**
5.1 Semiconductor fab as a complex job shop application and benchmark 137
5.2 Performance evaluation of time-constraint adherence 138
5.2.1 Performance metrics for the evaluation of Prediction Interval Quality . . 139
5.2.2 Performance metrics for the binary classification evaluation 140
5.3 Evaluation of Foresighted Digital Twin-based production control approach . . . 140
5.3.1 Evaluation of the simulation model prediction 141
5.3.2 Evaluation of binary classification 142
5.4 Evaluation of transitional model based production control approach 143

5.4.1 Evaluation of the influence of multi-variate prediction intervals 145
5.4.2 Evaluation of binary classification 146
5.5 Summary of evaluation and computational results 149

6 Discussion and Outlook **151**
6.1 Discussion 151
6.2 Outlook and further considerations 155

7 Conclusion **157**

8 List of own publications **158**

9 References **163**

List of Figures **188**

List of Tables **192**

Appendices **X**
A1 Additional data analysis of the real-world semiconductor manufacturing dataset X
A2 Additional experimental evaluation of transition time prediction XII
A3 Additional experimental evaluation of time-constraint adherence prediction . . . XIV

Abbreviations

Abbreviation	Denotation
AE	Absolute Error
AI	Artificial Intelligence
AR	Autoregressive (model)
ARIMA	Autoregressive Integrated Moving Average (model)
ARMA	Autoregressive Moving Average (model)
CMP	Chemical-mechanical polishing
CVD	Chemical Vapor Deposition
DES	Discrete Event Simulatin
DGP	Data Generation Process
DUV	Deep ultraviolet (light)
EUV	Extreme ultarviolet (light)
FN	False Negative
FOUP	Front Opening Unified Pod
IC	Integrated Circuit
IIoT	Industrial Internet of Things
IoT	Internet of Things
KB	Knowledge Base
KG	Knowledge Graph
MA	Moving Average (model)
MAE	Mean Absolute Error
MDP	Markov Decision Process
MILP	Mixed Integer Linear Program
ML	Machine Learning
MPIW	Mean Prediction Interval width
MSE	Mean Squared Error
MTBF	Mean Time Between Failures
MTTR	Mean Time To Repair
OWL	Web Ontology Language
PICP	Prediction Interval Coverage Probability
PLC	Programmable Logic Controller
PPC	Production Planning and Control
PVD	Physical Vapor Deposition
RAE	Relative Absolute Error
RAM	Random Access Memory

RDF	Resource Description Framework
Si	Silicon
SiO2	Silicon dioxide
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured query language
TN	True Negative
TP	True Positive
UQ	Uncertainty Quantification
UV	Ultraviolet (light)
WIP	Work in progress (inventory)
XML	Extensible Markup Language

Formula symbols

Formula symbol	Name	Unit
A	Set of all axioms (in Ontology)	-
A^{RL}	Set of actions (in RL)	-
a_t	Action at time t (in RL) with $a_t \in A^{RL}$	-
A^*	Constraint parameters (in linear optimization)	-
α	Coverage (for prediction intervals)	-
B	Set of non-taxonomic relations (in Ontology) with $B \subset R$	-
b^*	Inequality parameter (in linear optimization)	-
C	Set of all concepts (in Ontology)	-
c_t	Cell state at timestep t (in RNN)	-
c^T	Cost matrix (in linear optimization)	-
d	Order of differencing for ARIMA model	-
d^l	Lower bound of prediction interval	-
d^u	Upper bound of prediction interval	-
E	List of Equipment	-
e	Equipment with $e \in E$	-
e_i	Prediction error	-
ϵ_t	Error term at timestep t (in Time Series)	-
H	Set of taxonomic relations (in Ontology) with $H \subset R$	-
h	Number of hidden layers (in NN)	-
h_t	Output at timestep t (in RNN)	-
L	Inventory (average)	-
$L_{winkler}$	Winkler loss function	-
$L_{one-sided}$	One-sided winkler loss function	-
l	Lot	-
l_l	Lower bound in KS Test	-
l_u	Upper bound in KS Test	-
λ	Throughput (average)	-
λ_l	Prediction interval width trade-off parameter	-
λ_u	Prediction interval coverage trade-off parameter	-
μ	Assumed mean of a distribution	-
$O_{l,r}$	Operation number r of lot l	-
Ont	Ontology	-

Formula symbol	Name	Unit
ω_{ij}	Weights of connection between neurons i and j (in NN)	-
$P(x)$	Probability of event x	-
p	Order of Autoregressive model	-
π_{RL}	Policy (in RL)	-
ϕ_{t-i}	Autoregressive parameter for input $t-i$ (in Time Series Models)	-
q	Order of Moving Average model	-
R	Set of all assertions (in Ontology)	-
r_t	Reward at timestep t (in RL)	-
S^{RL}	Set of states (in RL)	-
s	Sample variance (for prediction intervals)	-
s_t	State at time t with $s_t \in S$	-
σ	Standard deviation	-
Top	Hierarchical highest concept (in Ontology)	-
t	Timestep	-
$t_{1-\alpha}$	Value of student's t-distribution at coverage $1-\alpha$	-
θ_{t-i}	Moving Average parameter for input $t-i$ (in Time Series Model)	-
$Var(e)$	Variance of error e	-
v	Loop (in RNN)	-
W	Waiting time (average)	-
X	Time Series $X = (X_1, X_2, ...)$	-
$\bar{X}$	Estimate for X	-
x	Variable	-
x_j	Input data point j (in NN)	-
(x_k, y_k)	Dataset in machine learning for $k \in \{1, 2, ..., N\}$	-
$\hat{y}_i$	Output data point i (in NN)	-
$\hat{y}$	Estimate of y (in prediction intervals)	-

1 Introduction

Technological leaps forward heavily increase technological complexity in manufacturing. In wake of rising, progressively fluctuating market demand and increasing changes in customers' requirements manufacturing's competitiveness is tested along these trends. Individualization fosters shrinking batch sizes (Gu & Koren 2018). High quality standards and increasing technological capabilities enforce more sophisticated machinery (Schmitt et al. 2012). As a result, operations become increasingly convoluted (Mönch & Fowler & Mason 2013) where technological progress tightens the space for operating manufacturing competitively.

Digitalization offers technological progress poised to enable the handling of greater organizational complexity in manufacturing. A transition towards smart devices and intelligent interconnections is fueling an ever-increasing demand for semiconductors to gather, process and store data (Mönch & Fowler & Dauzère-Pérès, et al. 2009). Their application in manufacturing gave birth to the fourth industrial revolution to enable tapping into great potential for operations improvement and intelligent automation (Lasi et al. 2014). Industry 4.0 draws from compendious data collections from production systems in real-time as well as their semantic enrichment (Abramovici et al. 2016). The availability of exhaustive general and real-time state information about products, machines and internal logistics empowers system optimization and transparency. Available real-time factory data is the sine qua non for operations management to tackle tightening, stringent requirements put on manufacturing (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Hence, progressive algorithms play a decisive role in permitting intelligent production control.

Artificial intelligence is over and over seen as a complement to human intelligence in decision making enabling smart manufacturing (Arinez et al. 2020). Machine Learning (ML), as a data-driven artificial intelligence, is the main driver of its recent prominence with examples ranging from superior chess or Go playing (Silver et al. 2018) and self-driving cars (Rao & Frtunikj 2018) to multivariate generative content creation (Wolf et al. 2020). The recent gains in prominence of Machine Learning, demanded by a highly competitive environment, is fueled by increasingly potent hardware, ever more powerful, freely available software, expanding data availability and improving algorithms (Wuest et al. 2016). Applications extend more and more beyond well described systems and datasets; however, the uncertainty within systems and models is less frequently regarded. Such system inherent uncertainty and complexity prevails in real-world production control. Thus, there is a large potential for improving operations management and enabling effective production control in complex production systems.

1.1 Motivation

Industrial manufacturing developed towards highly coordinated manufacturing systems. Specialization of equipment and personnel to performing individual tasks at a high quality and high speed enabled economies of scale (Helpman 1981). Convoluted material flow, more complex processes and interactions aggravated the need for an effective and efficient organization (Ueda et al. 2002). Production Control ensures cost-effective and stable operations (Czumanski & Lödding 2016).

Modern, complex job shops incorporate different lot sizes, inconsistent process times, frequent failure events, flexible and recurrent material flow as well as time constraints (Waschneck et al. 2016). These complexities aggravate the traditional production control tasks order release, sequencing and capacity planning among others. Increasingly frequent decisions and adaptations to the real-time circumstances in the production system become the norm (Altenmüller et al. 2020). Irrespective of the concrete production control objective preference in a modern production system waste must be eliminated. Therefore, the production control system must ensure no value creation opportunity is squandered and no value scrapped.

Traditionally, production planning and control divides the associated tasks along timely dimensions, long-term planning and short-term control (Mönch & Fowler & Mason 2013). Stochastic uncertainty decreases with decreasing size of regarded time intervals. Systematic uncertainty, in contrast, increases as operations convolutions are amplified. The latter; thus, contributes the major source of concern for production control. Increasing data availability and the need to cope with ever growing uncertainty provide a suitable environment for adaptive production control (Monostori et al. 2004).

More complex systems yield more sophisticated algorithms which can improve production control for less complex systems alike. Complex job shops constitute the most complex production systems (Waschneck et al. 2016). This complexity requires more adaptive, learning-based systems for production control as traditional deterministic planning is insufficient. Static, rule-based industry standards oversimplify complex problems (Nickel et al. 2014). Alternatively, humans are often performing complex decision making tasks. However, human intelligence is error prone, inconsistent in round the clock production, incapable of large scale data processing and expensive (Wegener et al. 2021). Automated adaptive, intelligent production control can alleviate these shortcomings.

Learning-based approaches can be both adaptive and intelligent by altering the behavior based on environmental conditions (Russell & Norvig 2021). Machine Learning (ML) is the most prominent example herein, capable of large-scale data processing. However, ML lacks explainability and the capability to effectively deal with varying epistemic uncertainty (Kuhnle

et al. 2022). For adaptive, intelligent production control of real-world complex job shops novel forms become necessary. Only few applications can be found in literature as real-time data-driven production control approaches in complex job shops are not yet common and have not accomplished the full potential for industrial application.

1.2 Problem statement

Augmenting production control towards intelligent production is applicable to a wide range of industries in the manufacturing sector. The present thesis is in particular motivated by the semiconductor industry as a manifestation of complex job shops. For application, however, this industry is superseded by any manufacturing industry that exhibits similar characteristics.

The electronic industry has become one of the largest industries world-wide as semiconductors are essential components in all industries nowadays (Valet et al. 2022). Its importance and growth is fueled by the global megatrends of Artificial Intelligence (AI), Industrial Internet of Things (IoT), smart products and the software-defined characteristics of modern systems. Embedding circuits into new product generations is critical for dealing with climate change and resource scarcity (Uçar et al. 2020). Hence, semiconductors act as enablers for achieving energy efficiency, individual mobility, security, the establishment of Industrial Internet of Things (IIoT) and the application of AI. This has led to fast innovation cycles in semiconductor manufacturing and thus its technology-intensive and capital-intensive nature (Mönch & Fowler & Dauzère-Pérès, et al. 2011).

Semiconductor manufacturing equipment accounts for a major cost driver (Hong et al. 2023) so that opportunistically utilizing capacities and avoiding the production of faulty products is paramount (Valet et al. 2022). Therefore, semiconductor fabrication plants, in short fabs, operate 24 hours a day on any day. To minimize coordination effort and setup times wafers containing Integrated Circuit (IC) chips are packetized into lots of 25 or 50 (Ziarnetzky et al. 2017). The technological intensity is manifested in the presence of forming and cutting processes, electro-physical and chemical processes, abrasive processes, surface engineering and metrology augmented by complex machinery and production system organization as well as highly specialized semiconductor design (Mönch & Fowler & Mason 2013). Beyond individual technological complexity the major challenge for semiconductor manufacturing lies in the coordination of this complex job shop (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Each wafer, a thin slice of semiconductor material, requires up to 800 processing steps each between a few minutes and several hours (Ziarnetzky et al. 2017). Recurrent visits to many machines are necessary to transform semiconductor material into an IC. To achieve economies of scale ICs are pooled onto a single wafer. Fabs operate on the verge of the physically, technologically possible resulting in only a share of the produced ICs being fully

usable (Mönch & Fowler & Dauzère-Pérès, et al. 2009). Keeping this share, called yield, high is decisive. In addition to processing or design errors a major yield loss consists of contaminated wafers through native oxidation, crystal formation, ion migration or dust deposition (Lima et al. 2021). These impurities pollute the surface and inhibit the designed electrical flow. Thus, semiconductor manufacturing equipment in fabs is operated in clean rooms to reduce contamination (Klemmt & Mönch 2012). Nevertheless, between many process steps wafers can only remain unprocessed for several hours otherwise yield is reduced as the wafers can often not be recovered and have to be scrapped (Altenmüller et al. 2020). Hence, adhering to these time-constraints is of utmost importance (Arima et al. 2015).

Controlling the production under time constraints, re-entrant and non-linear material flow given varying process times is challenging (Wang & Srivathsan, et al. 2018). System behavior is time transient as the overall time wafers spend within a fab can be up to several weeks and months (Mönch & Fowler & Mason 2013). In this volatile environment production control is amplified by human operators that deal with time-contraints (Lima et al. 2019). This requires additional manual effort, is error-prone, inconsistent and neglects optimization opportunities. Additionally, the process is time-consuming and stressful for operators (Lima et al. 2017a). Establishing an intelligent production control in this complex job, that deals with time-constraints, can alleviate the current shortcomings. Making use of the real-time fab data can enable overcoming traditional, rule-based approaches that are too rigid (Altenmüller et al. 2020). Additionally, reducing these wasteful activities not only contributes to monetary business objectives, operator well-being, but also to sustainability and the achievement of net zero climate goals as wafer fabrication is energy intensive (May & Behnen, et al. 2021).

1.3 Research hypothesis

The above outlined challenges and problem statement can be addressed by an overarching research goal as follows:

> *Intelligent production control for time-constrained complex job shops, based on real-time data.*

Modeling, implementing and evaluating such an intelligent production control to avoid time-constraint violations is focused. A semiconductor fab is chosen as a real-world use case that exemplarily demonstrates the approach and its feasibility. Therefore, as a prerequisite, this research aims at the implicit proposition that any given production system is describable by a semantically structured data set in form of a knowledge graph that contains all information relevant to delineate a static and dynamic representation. Given this static knowledge graph a digital replica of the regarded system can be created. With this digital replica's behavior akin

to the real system's scarce data can be augmented and behavioral predictions performed. Based on this semantically structured data and system replication human effort for modeling and evaluation shall be minimized without compromise in performance or scope.

The research agenda deduction is based on the analyzed literature so that the following five research questions will be answered within this thesis.

1. How to use static and dynamic knowledge graph based production system replicas to support production planning and control with time-constraints?
2. How to use real-world real time data to avoid time-constraint violations with a data-based approach for production control for complex job shops?
3. How to enrich and extend machine learning algorithms to accurately capture the aleatoric and epistemic uncertainty in large-scale complex job shops when predicting time-constraint adherence?
4. How to use long-term and real-time knowledge acquired within a factory to holistically reduce time-constraint violations with intelligent production control?
5. How does the learning-based intelligent production control for complex job shop perform in ensuring time-constraint adherence in a real-world setting?

1.4 Structure of this work

This work is structured as follows. Within the first part of the thesis the problem and its setting are described. Therefore, production planning and control and its application in complex job shops and the semiconductor industry is introduced in Chapter 2. Different, quantitative approaches are compared and the prerequisites for simulation-based and intelligent, learning-based approaches are laid out. A comprehensive literature review is presented in Chapter 3 and the research deficit is deduced. Chapter 4 proposes the developed methodology for intelligent production control for time-constrained complex job shops in three steps. First, the problem scope is specified and analyzed. Second, the modeling approach is expound. These constitute the bedrock of the proposed approach that is described. In Chapter 5 the approach is evaluated in a real-world system to assess its performance and applicability. The computational results compare the approach to the state-of-the-art and deduce requirements for real-world application. An outlook follows the discussion of the approach in Chapter 6. Ultimately, Chapter 7 presents a summary of the entire work.

2 Fundamentals

In the following chapter fundamentals that this thesis is built upon are introduced. The structure and interrelations are visualized in Figure 2.1. First, in Section 2.1 semiconductor manufacturing, the manifestation of a complex job shop, is introduced. Both a sound technical understanding and its influences on operating such a semiconductor manufacturing factory are discussed. Time-constraints as the major hurdle in effectively controlling this complex job shops are defined. As the main focus of this thesis rests on introducing novel, intelligent production control for complex job shops, the production planning and control framework is explained in Section 2.2. The complexities arising from complex job shops, traditional ways of implementing production control as well as requirements for embedding a novel production control are discussed. Nowadays, production control is often based on data and quantitative decision making. Section 2.3 introduces quantitative optimization as a tool suitable for decision making through effectively handling large amounts of data. Traditional approaches from mathematical optimization, heuristics towards artificial intelligence based knowledge management and machine learning are introduced. The advantages and limitations are discussed to conclude this section. Building on the real production a digital twin as its virtual counterpart is introduced in Section 2.4. As dealing with time series data is frequently necessary in modern manufacturing systems and particular complex job shops time series modeling as a Machine Learning approach is presented in Section 2.3 as well. Classical models and machine learning models are introduced and the applicability and technical realization of prediction intervals presented. As during the course of this thesis prediction intervals have to be extended to one-sided prediction intervals their formal derivation is proved. All in all, the foundations for understanding the state-of-the-art and own approach are found in this chapter.

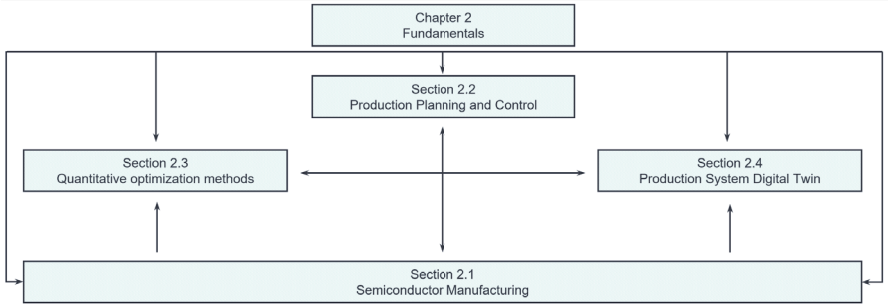

Figure 2.1: Structure of the Fundamentals chapter with relations between sections.

2.1 Semiconductor Manufacturing

In 1958 Jack Kilby invented the integrated circuit (IC) (Kilby 2000). These digital microchips are widely used in many products: logic chips for computing based on logic gates, memory chips to store data based on NAND gates and field-programmable gate arrays (FPGAs) that can be programmed after manufacturing (Pe 2018). Analog microchips are similarly common: radio frequency circuits and mixed-signal ICs such as analog(A)/digital(D) or D/A converters (Pe 2018).

Manufacturing of these ICs is known as semiconductor manufacturing as nowadays ICs are made of semiconductor material. Its conductivity can be altered by crystal structure impurities (Mönch & Fowler & Mason 2013). This modification of the electrical properties enables a plethora of applications as conductivity can depend on temperature, light or electrical fields (Pe 2018). Jack Kilby's first IC was made of the semiconductor germanium (Kilby 2000). Today, silicium is widely used to manufacture ICs (Mönch & Fowler & Dauzère-Pérès, et al. 2011). In general, any semiconductor is in principle a feasible material, so that use-case specialized the material has to be selected to favor strong currents or low energy consumption

IC design and fabrication can be decomposed into several steps as illustrated by Figure 2.2. First, chips and their function are designed and simulatively tested based on a chip development plan (Pe 2018). Secondly, the physical design process is used to produce photomasks which are the IC blueprints (Xiao 2012). In the third step in the front-end semiconductor manufacturing wafers are fabricated and tested (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Lastly, the wafers are cut and packed for final testing during the back-end semiconductor manufacturing (Pe 2018). The latter of this thesis will focus on front-end manufacturing, a complex job shop that decisively influences product quality and manufacturer competitiveness. As a semiconductor front end fab is the most complex job shop currently known dealing with this system will allow transferability.

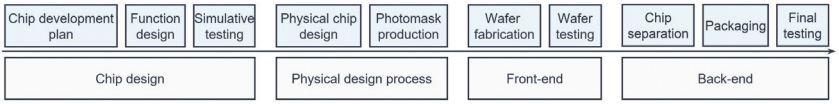

Figure 2.2: Semiconductor manufacturing overview. Adapted from Mönch & Fowler & Dauzère-Pérès, et al. (2011) and Pe (2018).

In the following, Section 2.1.1 introduces semiconductor technology followed by a description of the front-end in Section 2.1.2. Time-constraints as the major problem in complex job shop and semiconductor production control is introduced in Section 2.1.3. Section 2.1.4 summarizes semiconductor manufacturing requirements.

2.1.1 Semiconductor fabrication technology

Transistors form the basis of modern microchips which can contain several billion transistors (Pe 2018). A transistor can mimic binary behavior by controlling electrical flow interpreted as on/off or 1/0. Using the conditional conductivity of semiconductors, such as Silicon (Si), the passing or not passing of electrical currents is realized. As shown in Figure 2.3 a transistor is realized by controlling the electrical flow between source and drain through a gate (Pe 2018). Source and drain are a part of the silicon wafer that is doped through ion implantation in the silicon crystal lattice. Doping the silicon with impurities of five valence electron elements (such as phosphorus, arsenic or antimony) creates so called n-semiconductors as shown in Figure 2.3. Four of these electrons each connect to a silicon atom in the crystal structure around leaving one free electron. This free electron easily jumps into the conduction band and enables electric conductivity. Vice versa, three-valent elements (such as boron, aluminum or indium) create a free hole in the valence band as shown in Figure 2.3. Controlling this deliberate contamination precisely can realize tiny conductive paths in the silicon (Mönch & Fowler & Mason 2013). In a transistor the gate is insulated from source and drain as the gate voltage controls the electric flow between source and drain. If the gate voltage is positive electrons in the body are attracted to the surface, creating a conductive n-channel connecting source and drain. This acts as a 1 or on in a binary entity (Pe 2018) and is, thus, the building block of binary operations featured by microchips. Transistors are hence among the most common man-made, manufactured structures.

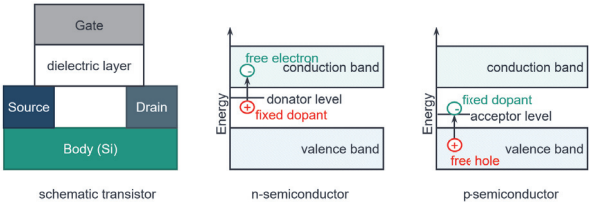

Figure 2.3: Transistor and comparison of n- and p-semiconductors based on Pe (2018).

Key to the success of ICs is their planar layer by layer manufacturability (Mönch & Fowler & Mason 2013). As visible in Figure 2.3 a source and drain are built on top of a silicon body nowadays made from ultra pure silicon wafers. Source and drain are manufactured through carefully n-doping the semiconductor in a first layer. Secondly, a dielectric layer that isolates can be made of thin oxide. Then, the gate can be added in another layer. Nowadays ICs consist of multiple layers and oftentimes three dimensional transistors. (Pe 2018)

This layerwise IC manufacturing operates on a pure crystalline silicone wafer (Chen & Tosello, et al. 2022). Continuing the example from Figure 2.3 and as a basis for n-channel field-effect transistors the silicon is p-doped. Figure 2.4 illustrates the process outlined in the following. The pure silicon wafer is then prepared by oxidation and diffusion to create a silicon dioxide ($SiO2$) top layer. In the next step photoresist is deposited. Lithography, often described as the holy grail of semiconductor manufacturing, exposes the photoresist to ultraviolet (UV) or deep ultraviolet (DUV) light filtered by a mask (Graff et al. 2023). The remaining photoresist protects underlying layers from subsequent etching. After photoresist removal the ion implantation dopes the silicon to create n-semiconductor (or p-semiconductor). Lastly, various removal process planarize the wafer and prepare the next layer. In the following the individual process steps are briefly introduced as each wafer passes through up to 1000 of these processing steps and hundreds of associated machines (Mönch & Fowler & Dauzère-Pérès, et al. 2011).

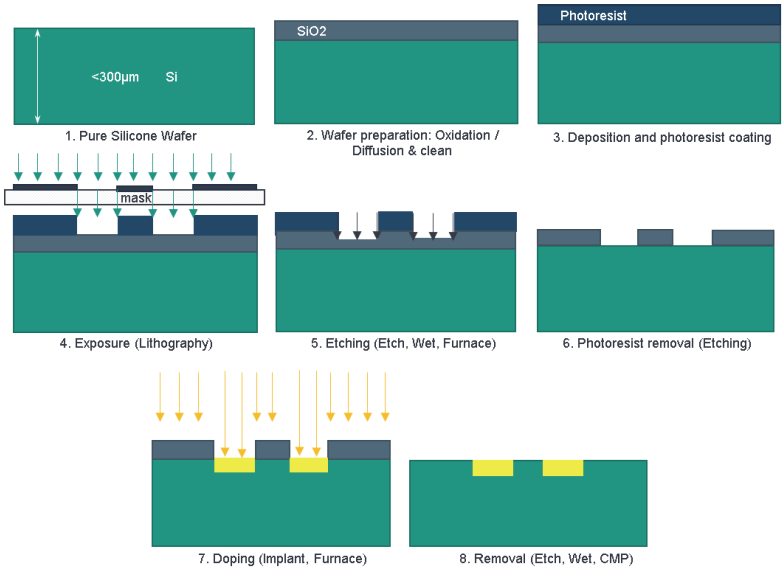

Figure 2.4: IC manufacturing steps. Adapted from Pe (2018) and Mönch & Fowler & Dauzère-Pérès, et al. (2011).

2.1.1.1 Silicon wafer manufacturing

A semiconductor wafer, or wafer, is a ultra-thin disc of semiconductor material measuring up to $300\mu m$ (Pe 2018). Currently, silicon is predominantly used and wafer diameters have

historically increased from recently $200mm$ to nowadays $300mm$ (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Greater wafer diameters increase the usable share of the wafer surface as the circumventing edge passes through relatively less ICs. This increases productivity. At the same time transistor sizes have been shrinked, according to Moore's law, to amplify the ICs computing power and decrease electricity usage (Mönch & Fowler & Mason 2013). This miniaturization of transistors to only a few nanometers (Pe 2018) necessitates higher silicon purity of wafers (Chen & Tosello, et al. 2022). Float-zone methods provide the highest quality of silicone crystal. By avoiding the contamination from a quartz crucible melt in Czorchalski methods impurities are minimized (Chen & Tosello, et al. 2022). This mono-crystalline silicon is then cut with high precision into single wafers (Pe 2018). Due to this complex expensive processes even blank silicon wafers contain a high inherent value. Therefore, contaminating or scrapping wafers in semiconductor manufacturing has to be avoided (Mönch & Fowler & Mason 2013).

2.1.1.2 Oxidation and Diffusion

After cleaning and wafer preparation oxidation is used to develop a $SiO2$, silicon dioxide, layer on the wafer (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Thermal oxidation, diffusing an oxidizing agent at high temperature into the wafer, is widely used (Pe 2018). High temperatures facilitate quick oxidation, so that furnaces are used. For operating a wafer fab furnaces typically have large oxidation chambers that can contain multiple wafers at once (Mönch & Fowler & Mason 2013). Depending on the oxidizing agent used, wet (H_2O) and dry (O_2) oxidation are distinguished, leading to different oxide density, dielectric strength and oxide layer thickness (Xiao 2012). Thus, from a fab perspective, the need to use the right oxidation equipment and control material flow and furnace fill levels arises (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Furthermore, these furnaces with high temperatures require ultra clean wafers as thermal treatment of contaminants can lead to short circuits. Therefore, wafers are only handled and treated in a clean room to minimize contamination (Sun et al. 2005).

2.1.1.3 Deposition

Deposition aims at depositioning a thin layer on the wafer (Mönch & Fowler & Mason 2013). Dielectric or metal layers are deposited with physical vapor deposition (PVD) and chemical vapor deposition (CVD). In ultra clean, high vacuum environments PVD evaporates source material from a solid or liquid phase. This gase phase is then transported and deposited on the wafer as a condensed phase (Pe 2018). Most notable processes include molecular beam epitaxy, enabling precisely controlled deposition areas, evaporating and sputtering, depositing striked out ions. CVP on the other hand, uses gases for deposition and operates with different techniques from atmospheric pressure till low pressure (Xiao 2012). Aside from

different pressures plasma enhanced CVD can provide high speed deposition whereas atomic layer deposition enables three dimensional structure deposition (Pe 2018). Wiring can be achieved by metalization (Mönch & Fowler & Dauzère-Pérès, et al. 2011) which can deposit aluminum or copper. Again, the correct process equipment and material is paramount not to scrap the wafers during operation or due to contamination (Mönch & Fowler & Mason 2013).

2.1.1.4 Photolithography

Before the photoresist coating the wafer has to be cleaned (Pe 2018). Photoresist can be negative, i.e. areas undergoing light exposure become harder to dissolve, or positive, each using different chemical structure, exposure and having different resolution capabilities. Core of lithography is exposing the photoresist coated wafer to high energy radiation filtered through a mask (Mönch & Fowler & Mason 2013). The complexity of lithography arises from perfectly aligning the mask with the layers under the photoresist and controlling the optics for correct focus and exposure (Graff et al. 2023). Decreasing transistor size requires shorter wavelengths leading to nowadays DUV or extreme ultraviolet (EUV) systems capable of exposing transistors of only a few nanometers in size. Physical and chemical defects, for instance from particle interference or refraction, should be avoided. Nowadays, run to run control compensates based on previous runs and occasional test wafers (Graff et al. 2023). Hence, processing orders and required test wafers can hardly be predicted. Additionally, contamination can not only damage individual ICs on the wafer, so that scrap should be avoided. Lithrography tools are among the by far most expensive equipment and thus often present a bottleneck in semiconductor manufacturing (Mönch & Fowler & Mason 2013).

2.1.1.5 Etching

Etching aims at removing whole layers or material in photoresist coated layers on the wafer's surface (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Again, wet and dry types can be distinguished (Mönch & Fowler & Mason 2013). While dry etching applies gas wet etching is chemical bath based. The latter is often implemented in batch etching which soaks and lifts wafers into etch and cleaning baths. These etch chemicals need to be renewed regularly and uniformity is hardly achieved (Pe 2018). Alternatively, spray etching implements a lot size one wafer etching based on gas. For all processes temperatures and chemicals vary heavily. Depending on the material to be etched setup and cleaning times can be high (Xiao 2012). Photoresist can also be removed by plasma ashing also called plasma etching. Alternative dry etch process include ion beam etching and reactive ion etching. All in all, etching is crucial

as underlying layers should not be damaged (Chang & Chang 2012).

2.1.1.6 Ion Implantation

Ion implantation dopes the unprotected, etched areas with positive or negative ions to create p- or n-semiconductors (Pe 2018). Electrically charged ions are directed from an ion source to the wafer's surface (Mönch & Fowler & Mason 2013). Photoresist blocks the ions from entering the silicon and altering its conductivity where it is not desired. Implant equipment works with high electricity and ion sources, and thus exhibits frequent required maintenance (Yang & Ke, et al. 2015). Furnace and etch operations can occur afterwards to remove the photoresist. Due to the ion beam, wafers have to be processed one by one. In general, four to eight times, or sometimes even more regularly, wafers visit implant equipment (Mönch & Fowler & Mason 2013).

2.1.1.7 Planarization

Planarization is performed with the removal of the siliconoxide layer to clean and level the wafer surface (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Chemical-mechanical polishing (CMP) smooths the surface with mechanical force and chemically acting slurry. A large pad is mechanically, abrasively polishing the smaller wafer that is soaked into a chemical (Pe 2018). However, this abrasive process can induce stress cracks, create particles or come with impurities. Each CMP can only polish a single wafer at a time which requires cleaning afterwards. Leveling the surface enables photolithography focus (Mönch & Fowler & Mason 2013).

2.1.1.8 Contamination and cleaning

Frequent cleaning of wavers is necessary, especially after wet etching, to clear the wafer from particles (Xiao 2012). Contamination is a crucial factor that renders chips or wafers unusable leading to scrap. In general, there are different types of contamination stemming from resist or solvent residuals, molecular contamination, ambient air, abrasion and humans in form of microscopic contamination and alkaline or metallic contamination (Pe 2018) as illustrated in Figure 2.5. Microscopic particles from ambient air, abrasive process, airborne clothing particles or inadequately filtered liquids attach to wafers and hinder correct exposure, implant or etching process shown in Figure 2.5. This can lead to ICs being surrendered unusable, and hence whole wafers that need to be scrapped if too contaminated (Perraudat et al. 2019). Embedding these particles within layers can likewise create unevenness or after CMP lead

to short circuits. Entitled molecular contamination based on its molecular size, photoresist and solvent residuals or oil mist accumulate on the wafer surface, diffuse into the wafer or hinder later adhesion of deposition. Such electric flow alterations similarly render wafers to scrap (Pe 2018). Alkaline or metallic contamination originates from ionic deposits from salty, human sweat, insufficiently deionized water or undesired sputtering of the manufacturing equipment. These contaminations stick to the wafer surface or enter the crystal structure capturing or releasing additional electrons. By affecting the electric behavior, in a similar vein, wafers potentially have to be scrapped.

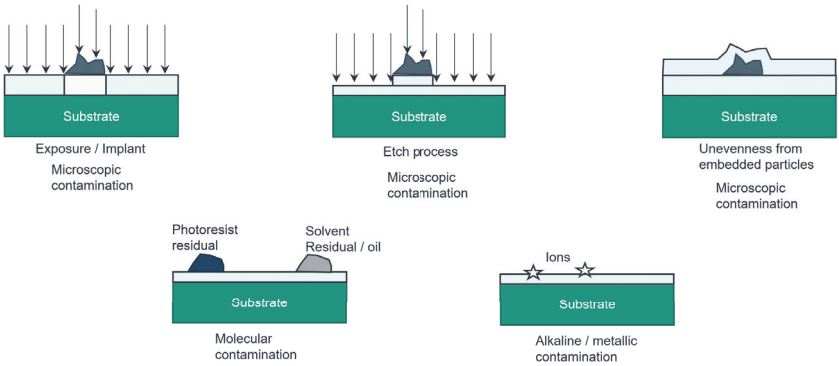

Figure 2.5: Frequent types of contamination in semiconductor manufacturing.

To address these risks, various cleaning techniques are introduced. For instance, ultrasonic baths stipulate particles and contaminations to be diluted in a cleanser. Additionally, nitrogen might blow off lightly bonded particles and high pressure cleaning can purge some contaminations. Solvents treat organic contaminations, however, solvents themselves contribute to molecular contamination. Ionic contaminations can be removed with deionized water and rotating brushes, but these accumulate particles and easily damage the wafer's surface. (Pe 2018) As these cleansing steps are often insufficient aggressive etching has to be used. However, this and further treatments often damage the wafer's surface or lead undesired oxide layer foundation (Xiao 2012). Therefore, precise cleaning sequences are required and as smaller structures hinder effective cleaning it cannot rescue all contaminated wafers. Hence, ensuring only minimal contamination through technical and organizational measures is of utmost importance (Maleck & Nieke & Bock & Pabst & Schulze, et al. 2019).

From a technical point of view, semiconductor fabs operate in clean room environments (Mönch & Fowler & Dauzère-Pérès, et al. 2011). There are several classes of cleanrooms, ranging from ambient air (400 million particles of size 5 microns in 1 m^3) to ultra clean

cleanrooms reaching only 10 particles of 0.1 microns. Within the cleanroom, freshly cleaned air is released at the ceiling and filtered through holes in the floor (Mönch & Fowler & Mason 2013). Doing so creates a vertical laminar flow that directs particles to the filter (Pe 2018). Nevertheless, better cleanroom classes are more expensive and hardly realizable in a big semiconductor manufacturing shopfloor. Therefore, Front Opening Unified Pods (FOUPs) separate the wafers in micro-environments of highest cleanroom standards (Mönch & Fowler & Mason 2013) while the transport and shopfloor areas only achieve medium cleanroom classes, thus, saving costs and environmental damage. FOUPs are often referred to as lots as each FOUP contains typically a lot of 25 individual wafers (Xiao 2012). Processing these lots at once, hence, is imposed to ensure proper material flow (Mönch & Fowler & Dauzère-Pérès, et al. 2011). These lots are tracked flowing through the semiconductor fab to control contamination levels and process control. Thus, manufacturing processes typically range from sequentially handling all wafers in a lot to parallelization (Xiao 2012). Note, that depending on the customer order size, not all lots have to consist of full 25 wafers (Mönch & Fowler & Mason 2013).

2.1.2 Frond-end wafer fabrication

The main value creation, operating the capital intensive and specialized manufacturing equipment introduced in Section 2.1.1, is performed during IC fabrication on wafers as shown in Figure 2.6. In a consecutive steps defective wafers are separated during testing complementing the front-end (Xiao 2012). Back-end is often geographically disconnected and performed in lower wage countries as a large degree of manual labor is involved. Thus, semiconductor manufacturing takes place in an extensive global production network as defined by Lanza et al. (2019). During the back-end operation scrap is still abundant. Wafers that are not scrapped in the front-end still contain an unnegligable amount of non-functional ICs (Pe 2018). During cutting and packaging functional ICs might be destroyed (Xiao 2012). Thus, only after final testing the ICs can be delivered to customers. The percentage of microchips that meet their electrical specifications at the end is called yield (Mönch & Fowler & Mason 2013). As processing costs can only be recovered through selling function microchips, sustaining a high yield is decisive in semiconductor manufacturing (Mönch & Fowler & Dauzère-Pérès, et al. 2011).

Besides controlling costs a high yield is necessary to control the total time required for manufacturing (Pe 2018). The lower the yield the more wafers need to be manufactured to attain to the orders. Hence, by increasing the yield less processes need to be performed during manufacturing as wafer- and batch-wise process steps for unnecessary wafers are saved. On time delivery is particularly stressing as a wafer requires hundreds of process steps and several months to be manufactured (Mönch & Fowler & Mason 2013). Wafer fabrication in the

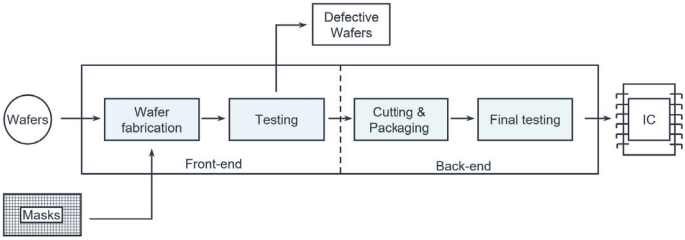

Figure 2.6: Wafer fabrication in a semiconductor fab. Adapted from Xiao (2012).

front-end makes up more than 75% of the total cycle time in semiconductor manufacturing (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Given the enormously high costs for semiconductor manufacturing equipment in wafer fabs the vast majority of manufacturing costs is incurred in the front-end (Mönch & Fowler & Dauzère-Pérès, et al. 2011). This triple of dominant scrap rate, cycle time and cost influence is fundamental for wafer fab importance. Thus, in the following the main focus of this thesis, as applied to semiconductor manufacturing, will remain on wafer fabrication in the front-end as a complex job shop.

A wafer fab's material flow is characterized by iterating through the base processes over and over again as illustrated in Figure 2.7. Each lot, and hence each wafer, passes 300 to 700 or sometimes more process steps (Mönch & Fowler & Mason 2013). This creates reentrant material flow as wafers are repeatedly passing through fab for each layer (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Moreover, auxiliary process and material flow must be controlled accordingly. This includes mask carrying reticles in lithography or chemicals in wet etching (Mönch & Fowler & Mason 2013). Additionally, human operators are required to successfully run a wafer fab in particular for inspection, maintenance and supporting material flow. Therefore, no full information about the material flow and lot positions are available. Lots can be stored in various first-in-first-out (FIFO), last-in-first-out (LIFO) or otherwise controlled buffers and storages (Mönch & Fowler & Dauzère-Pérès, et al. 2011; Pe 2018).

Processing times vary depending on the process step, processing equipment, previous process steps and wafers, batching and wafer requirements (Xiao 2012), where batching refers to collective processing of multiple lots at a time. Hence, processing times are stochastic and vary between minutes and 12 hours or more (Mönch & Fowler & Mason 2013). Combined with reentrant flow and batch processed wafers that are offloaded to single wafer operating machines, long queues become common (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Reducing inventory to shorten queues and cycle times is not economical. Capital intensive

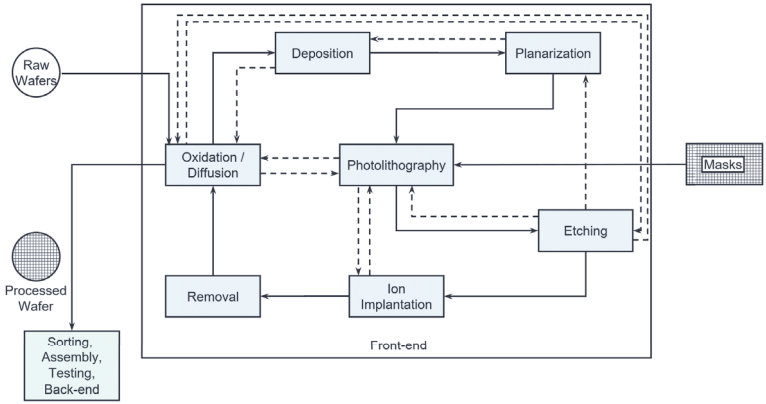

Figure 2.7: Wafer flowchart for a wafer fab. Adapted from Mönch & Fowler & Dauzère-Pérès, et al. (2011).

semiconductor manufacturing equipment must be fully utilized. Machine dedications and hot jobs aggravate the situation (Mönch & Fowler & Mason 2013) as explained in the following.

Machine dedications describe the fact that seemingly identical semiconductor manufacturing equipment is not capable to perform certain processes. Unless lengthy and expensive preparation and calibration is performed a single machine does not have the dedication to follow specific process patterns (Mönch & Fowler & Dauzère-Pérès, et al. 2011). If omitted, the process quality dips and wafers typically have to be scrapped. Hence, the order of lot processing is highly dependent on these dedications as illustrated in Figure 2.8. As equipment is operated on the verge of the physically possible equipment failures occur often and suddenly (Xiao 2012). Downtime can account for up to 40 % of the time (Mönch & Fowler & Mason 2013). Figure 2.8 shows the great influence on cycle times from machine failures and stringent machine dedications. Probabilistic failure length and occurrence as well as the time variance of machine dedications aggravate this problem (Pe 2018). Furthermore, sequence dependent setup times complicate cycle time estimations (Xiao 2012).

Hot jobs, which are close to their due date, are rushed through the wafer fab by jumping lines. Thus, they aggravate the congestion within the fab (Mönch & Fowler & Mason 2013). Additionally, so called engineering jobs are required as prototypes or to control processes (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Thus, preventive maintenance and engineering jobs further reduce processing capacity. Beyond these complexities time-constraints present the greatest obstacle for effectively and efficiently running a wafer fab (Klemmt & Mönch 2012). A time-constraint limits or in other words constrains the time a wafer can spend between any

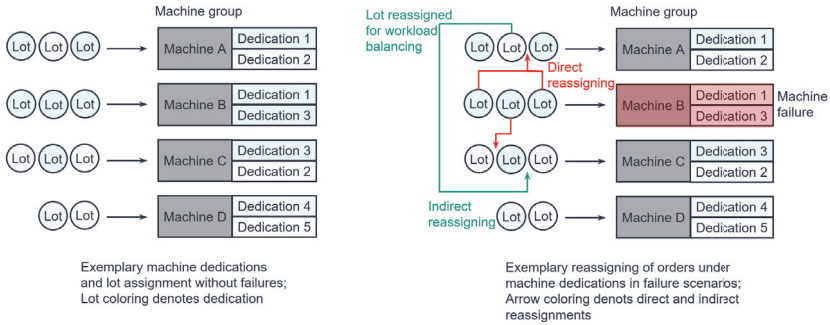

Figure 2.8: Machine dedications and machine failures influence cycle times and the processing order.

two processes before contamination prohibit further processing (Klemmt & Horn, et al. 2008). Thus, the transition time which measures the time a lot is in transition between two processes, is to be controlled to achieve time-constraint adherence. Rework is in general not permissible in cases of violated time-constraints leading to scrap (Mönch & Fowler & Mason 2013).

2.1.3 Time-constraints in Semiconductor Manufacturing

Strict process control is paramount in semiconductor manufacturing. In addition, non-process related influences need to be understood so that strict rules governing high yield can be imposed (Xiao 2012). Time-constraints severely distinguish semiconductor manufacturing from traditional manufacturing, as the time between process steps becomes limited to avoid contamination and deteriorating quality (Mönch & Fowler & Mason 2013). No universal term is used to describe time-constraints within and beyond wafer fabrication. The definition is that time-constraints regard two process operations, that are not necessarily consecutive, which are linked by a time limit that is specific for a wafer and, hence, lot (Klemmt & Mönch 2012). They arise from procedural restrictions when working with chemical and physical processes and concerns of avoiding contamination (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Therefore, only a subset of operation pairs are time-constrained for each wafer. Time-constraints differ between different wafers (Klemmt & Mönch 2012). Minimum and maximum time-constraints are known (Yugma et al. 2012). Adhering to minimum time-constraints is trivial if buffers are available as the respective wafer or lot can remain in the buffer until the minimum time has passed. Maximum time-constraint adherence, however, is difficult to achieve in real-world environments (Maleck & Weigert, et al. 2017). In general, time-constraints provide the greatest challenge for semiconductor production control as rework is often not permitted and lots must be scrapped (Lee 2020).

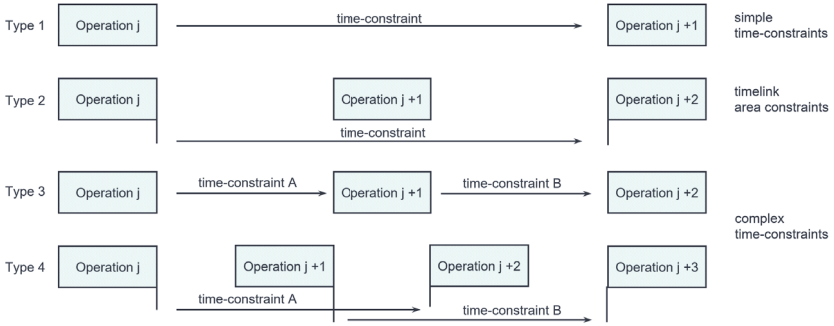

Figure 2.9: Time-constraint types limiting the transition time between two operations classified based on Klemmt & Mönch (2012) and Wang & Srivathsan, et al. (2018).

Commonly used terms include time-constraint, time constraint (Altenmüller et al. 2020; May & Maucher, et al. 2021) or time coupling constraint (Pirovano et al. 2020). These are often extended to waiting time constraint (Lee & Li 2017), queue time constraints (Ono et al. 2006) or similar terms in order to highlight the focus of individual studies. In concrete terms several types of time-constraints can be distinguished as illustrated by Figure 2.9. Consecutive operations that are linked through a time-constraint are often referred to as simple time-constraints (May & Maucher, et al. 2021). The term timelink areas for non-consecutive operations was introduced by Maleck & Weigert, et al. (2017). Three further types of time-constraints were introduced by Klemmt & Mönch (2012). These can be subsumed under the term complex time-constraints. Two consecutive simple time-constraints increase the complexity as the beginning of operation j directly limits operations $j+1$ and $j+2$ without leeway for shifting the central operation. Overlapping time-constraints in a similar vein limit the room for decision-making to ensure time-constraint adherence (Klemmt & Mönch 2012). Any combination of the above time-constraints is a complex time-constraint (Wang & Srivathsan, et al. 2018).

Another term frequently used is time constraint tunnel (Lima et al. 2021). The analogy of a tunnel very well suits the understanding of a time-constraint as products are in a tunnel as soon as the processing in the first operation has been started. Within this tunnel influence on the wafer or lot can only be exerted through controlling the material flow. However, physical transport can hardly be speed up and queuing positions are not freely interchangeable as several lots compete for the next processing slot due to time-constraint, due dates or process parametrization requirements (Mönch & Fowler & Mason 2013). Hence, controlling time-constraint adherence in wafer fabrication is a gate keeping decision (Maleck & Weigert, et al.

2017). In front of any equipment these gate keeping decisions release time-constrained or non time-constrained lots for processing. Due to the complexity of this gate keeping task and to facilitate the application of learned pattern and knowledge, nowadays, human operators often take these decisions (May & Maucher, et al. 2021).

To formally describe time-constraints, the following notation shall be used. The existence of a time-constraint is independent of the equipment $e \in E$ that is used. O describes a process operation on a semiconductor manufacturing equipment. Operations are wafer or lot specific expressed by O_l being an operation of lot l. A single wafer or lot l is described by the finite sequence of operations required to transform the pure wafer into ICs: $(O_{l1}, O_{l2}, O_{l3}, ..., O_{ln})$ with $n \in \mathbf{N}$.

Formally, a time-constraint, thus, has the definition of being two operations O_{lr}, O_{ls} where $r < s$ holds. O_{lr}, O_{ls} are inter-linked by a lot-specific (upper) time limit t_{lrs} (May & Behnen, et al. 2021). This time limit restricts the maximum time between the completion of operation O_{lr} and the start of operation O_{ls} for lot l (Klemmt & Mönch 2012). Additionally, multiple time constraints can be nested or or one time-constraint can be preceded by another, commonly known as complex time-constraints (Wang & Srivathsan, et al. 2018). Hence, complex time-constraints can be described based on their component-wise time-constraint disaggregation.

2.1.4 Summary: Semiconductor manufacturing complexities and requirements

In a nutshell, semiconductor manufacturing comes with process, equipment and externally infused complexities. Figure 2.10 outlines the greatest challenges and complexity drivers. While externally influenced complexities have less effects on short term decision making, their operational influence should not be overlooked. Semiconductor manufacturing has, thus, developed to a high-mix high-volume (Mönch & Fowler & Mason 2013) manufacturing environment. External influences, in particular fluctuating, sudden demand coupled with aggressive due dates, additionally lead to the status quo of operational excellence being the only path to success in this increasingly competitive environment (Pe 2018). Equipment related complexities additionally induce stochastic process times or sequence dependent setup times and quality from frequent breakdowns and machine dedications. This technological influence lead to yield becoming the single most important performance target in wafer fabrication (May & Behnen, et al. 2021). From an operational point of view system wide process related complexities pose the biggest challenge (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Convoluted, reentrant material flow, a mixture of one-piece flow, batching and further process types demand dynamic analyses. Stochastic equipment behavior aggravates this challenge and makes predictions hardly possible (Altenmüller et al. 2020). All these

complexities culminate in ensuring time-constraint adherence which, hence, remains the most challenging problem in operating a wafer fab (Maleck & Weigert, et al. 2017). Violating time-constraints directly decreases yield and delays deliveries due to the encountered scratch. Hence, controlling time-constraint adherence is of utmost importance.

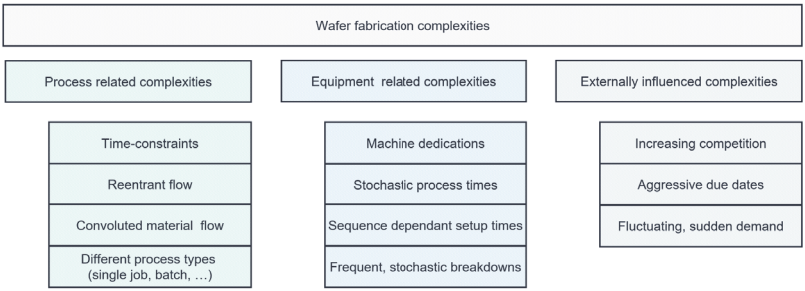

Figure 2.10: Complexities in wafer fabrication based on Mönch & Fowler & Mason (2013), Xiao (2012) and Valet et al. (2022).

To that end, the following requirements must be considered. First and foremost, time-constraint adherence must be ensured. Secondly, operational efficiency, in other words the utilization of capital intensive equipment, should not be compromised. Thirdly, the complexity in a semiconductor fab must be regarded dynamically and knowledge should be derived or used. By addressing these requirements, semiconductor manufacturing can become more efficient, more cost effective and more sustainable.

2.2 Production planning and control

Production is defined as the creation of products from various resources (Eversheim & Schuh 2013) that can involve both manufacturing and assembly. With increasingly complex manufacturing processes, products and an ever increasing demand for customized products production organization has evolved from an organization based on functionality to segmentation to strategic supply chains to finally global production networks (Wiendahl & ElMaraghy, et al. 2007). Cost-effectiveness and rapid reactions remain the key to produce successfully (Koren et al. 1999). All in all, production planning and control (PPC) nowadays aims to reduce the production company internal complexity and react rapidly and cost-effectively (Wiendahl & Breithaupt 1999). In a nutshell, PPC is the lever to reduce or control complexity in complex job shops and in generally complex manufacturing environments. The regarded semiconductor manufacturing fab exhibits these complexities as a prime example of a complex job shop. Therefore, in the following, PPC, its goals and approaches are introduced and discussed on a general perspective as well as the application to complex job shops.

2.2.1 Key Performance Indicators

Striving to successfully attain strategic goals is of utmost importance to withstand nowadays competitive manufacturing environment. Through monitoring and managing Key Performance Indicators (KPIs), that serve as an effective translation of business models, this success can be accomplished (Maté et al. 2012). KPIs are widely used in industry at different levels of decision making (May & Fang, et al. 2022). In PPC KPIs often serve as benchmarks and targets that have to be balanced and achieved (Wang & Liu 2013). Table 2.1 gives an overview of the typically most relevant performance measures that are captured in KPIs. They can be directly related to semiconductor manufacturing as capital intensive equipment should be utilized to recover high investment costs. Short waiting times are required to deliver on time and reduce inventory. In turn this is beneficial for time-constraint adherence (Kitamura et al. 2006). Violating time-constraints in semiconductor manufacturing and in general manufacturing inferior quality decreases yield and hence the operational performance. To achieve a good operational performance a high yield should be complemented with a high throughput leading to more quality assured products being manufactured at the same time.

Table 2.1: Most relevant KPIs as performance measures in production planning and control

Performance measure	**Unit**	**Description**
Utilization	$\%$	Percentage of time in which the object of study is utilized
Waiting time	$time$	Time that is not spend in value creation
Yield	$\%$	Rate of final products over produced that fulfills all quality requirements
Throughput	$\frac{products}{time}$	Number of products produced in a time period
Inventory	$products$	Number of products within the regarded system

Nonetheless, several of these performance measures and targets conflict with one another. The main goals in PPC, as identified by Wiendahl (1997), influence each other. As the above mentioned performance measures can be classified into these five categories they incorporate relationships such as that decreasing inventory comes at the expense of decreasing utilization. Similarly, decreasing waiting times requires decreasing inventory whereas throughput needs a balance of both. Thus, this problem folds into a multi-criteria optimization with different points in the Pareto Continuum to be achieved with different methods but ultimately a trade-off

decision has to be taken. Most famously, *Little's Law* connects the main KPIs of a discrete manufacturing production system (Little 2011) as follows:

$$L = \lambda W, \tag{2.1}$$

where L describes the average number of products in the system (inventory), λ describes the average flow through the system (throughput) and W denotes the average waiting time per product (Little 2011). The regarded system can vary from a single machine, multiple machines or an entire production system up to global production networks. Balancing the trade-off of these and further KPIs is the core task of PPC on different levels. Hence, recognizing the influences on other KPIs is of major importance when designing intelligent production control.

2.2.2 Production Planning and Control tasks

PPC is performed within a production company and plans ahead long-, mid- and short-term operations. Customers' and suppliers' requirements, however, are taken into account as PPC can "cut across company boundaries" (Wiendahl & ElMaraghy, et al. 2007) typically within a global production network. As a comprehensive value creation proposition PPC comprises activities from dispatching, scheduling, buying and maintaining machinery to organizing raw materials and human resources (Eversheim & Schuh 2013). Producing in time with the right quality at the right costs (Wiendahl & Breithaupt 1999) is paramount. Therefore, the production system, available capacities as well as capabilities, current orders and the expected demand are monitored through PPC. Through a company internal focused scope the production is managed as PPC controls the aforementioned factors. Individual sub-entities within a production company, such as sales, purchasing or storing, are lacking a holistic overview of the entire company so that coordination is preeminent. PPC thus ensures that orders are fulfilled and raw materials are sourced in a timely, qualitatively and cost-effectively manner (Wiendahl & Breithaupt 1999).

Coordination in PPC is executed through the control of material flow from raw materials to customer delivery, information flow between entities and financial flow (Wiendahl & ElMaraghy, et al. 2007) as shown in Figure 2.11. Financial planning governs the primary financial flows while PPC can allocate funds within a company through actions. Linking these entities and flows is the foremost goal of PPC to ensure a smooth production. Thereby, PPC deals with an ever-changing environment and a plethora of stochastic influences (Wiendahl & Breithaupt 1999). To contend with this uncertainty predictions and approximations are typically applied

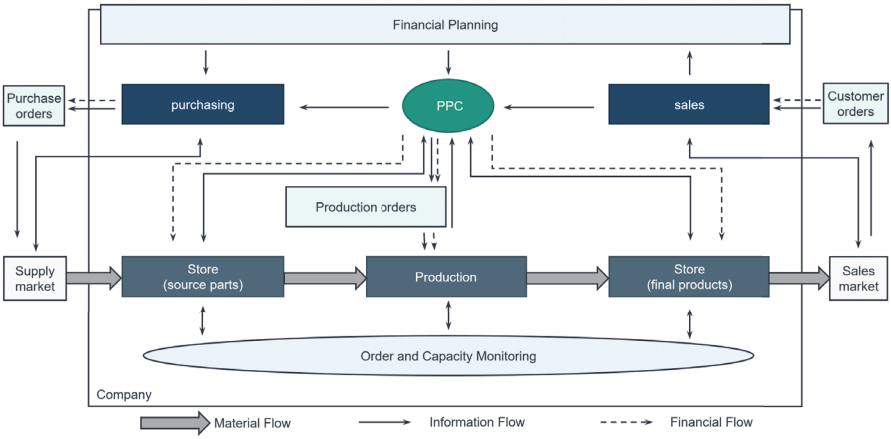

Figure 2.11: Production Planning and Control scope in a market economy, based on Wiendahl & ElMaraghy, et al. (2007)

through elaborate PPC software and potent computer systems (Wiendahl & ElMaraghy, et al. 2007).

Production Planning and Control (PPC) contains tasks that can be separated by the timely scope of their planning or control horizon and costs for reversing made decisions (Schuh & Wiendahl 1997). This leads to the three levels of long-term strategic decisions - implementing comprehensive company goals -, medium-term tactical decisions - planning product volumes, resources and aggregated inventory control - and short-term operational decisions - direct control of production resources and goods, including physical material flow (Nickel et al. 2014). Figure 2.12 unveils their interactions and the hierarchical structure. Decomposing the continuous PPC problem into hierarchical, individual problems enables the application of complex, yet solvable mathematical models, on each level (Nickel et al. 2014). Epistemic uncertainty increases in longer-term decisions. Aleatoric uncertainty is increasingly present in short-term decisions as the stochastic nature of events and their interdependence magnify the complexity (Wiendahl & Breithaupt 1999). Therefore, the complexity of PPC rises with the complexity of the regarded production system.

While the long-term targets PPC aims at include strategic goals and financial independence each level disaggregates these targets into Key Performance Indicators (KPIs) (Eversheim & Schuh 2013). This decouples the time invariance of decisions and avoids a hardly possible evaluation of long-term effects in short-term PPC tasks (Schuh & Wiendahl 1997). Therefore, KPIs are akin to the target measures for PPC that are to be achieved with the levers provided

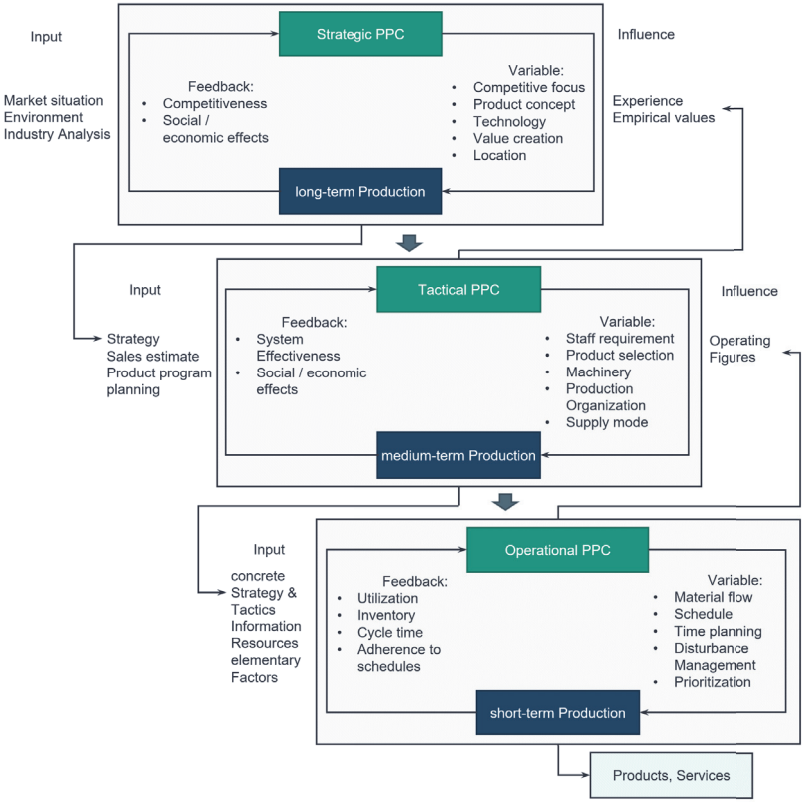

Figure 2.12: Levels of Production Planning and Control (PPC), based on Schuh & Wiendahl (1997); Mönch & Fowler & Mason (2013) and Nickel et al. (2014)

on each level as illustrated in Figure 2.12. Decoupling the decision making on individual levels is a major step for complexity reduction (Wiendahl & Scholtissek 1994). However, the influence of these complexities cannot be underestimated and must be regarded on all PPC decision making levels (Wiendahl & Scholtissek 1994). Figure 2.13 presents the most relevant traditional complexity drivers for PPC, which climax in operational PPC and short-term decision making with many degrees of freedom.

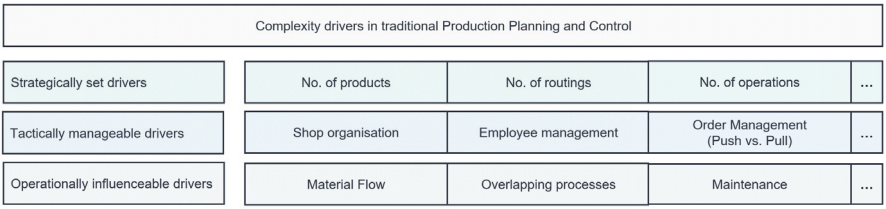

Figure 2.13: Traditional complexity drivers for PPC in manufacturing, based on Wiendahl & Scholtissek (1994)

Strategic PPC governs long-term decisions about product concepts, associated technologies (Schuh & Wiendahl 1997) and global production decisions on strategy, competitive niches or network footprint (Lanza et al. 2019). Hence, the decision horizon is in the order of months or years (Nickel et al. 2014) and uses empirical values obtained during past production, forecasts and predictions as well as experience. Thus, static optimization problems and intuition are often involved.

As opposed to tactical PPC which balances staff, machinery, supplier modes with product selection and required production targets (Nickel et al. 2014). The decisions are based on strict requirements such as production targets and intangible aspects such as strategy from the strategic PPC (Wiendahl & ElMaraghy, et al. 2007). One interesting problem here is aggregate production planning (May & Kiefer & Frey, et al. 2023) which controls workforce, machinery and inventory levels on a typically week-to-week basis. A clear input comes from operating figures from the shopfloor. These directly influence the values and expectations which are used in tactical PPC decision making which typically uses heuristics or linear static optimization approaches (Nickel et al. 2014).

Operational PPC, in contrast, directly controls the material flow, scheduling of jobs on machines and maintenance on a shopfloor level (Schuh & Wiendahl 1997). Due to this near real-time decision making it transforms strategic and tactic decisions within the space of feasible solutions given staff, machinery and inventory levels to concrete assignments on the production shopfloor (Nickel et al. 2014). Uncertainty about the times and costs diminish while

the uncertainty in general due to stochastic influence, e.g. breakdowns, increases (Mönch & Fowler & Mason 2013). A greater complexity in manufacturing directly relates to more complex PPC decision making (Wiendahl & Scholtissek 1994).

2.2.3 Complex Job Shops and Production Planning and Control

Increasing complexity in manufacturing is a major driver for changes in production paradigms (Wiendahl & Scholtissek 1994). As explained in Figure 2.15 function oriented job shops (Gabel & Riedmiller 2012) gave way to streamlined serial production. In order to cope with technological requirements and frequent changes flexible manufacturing systems (FMS) introduced free material flow in clustered connected equipment environments (Wurster et al. 2022). Easy reconfiguration based on a modular production system paradigm with standardized modules extended these to reconfigurable manufacturing systems (RMS) (Koren et al. 1999) and finally matrix production with a grid like structure (May & Schmidt, et al. 2021). Achieving this takt time independence, redundancy and loose material flow leads to a more job shop like structure. By drawing upon the previous concepts and recognizing the vast complexities along real-time decision making from setup to processing and material flow or time-constraints the term complex job shops was coined(Waschneck et al. 2016). Thus, the main different between matrix production and complex job shops lies in the operational complexity enforced in complex job shops that stem from convoluted material flow through reentrant jobs, different lot sizes as well highly variable process or setup capabilities and times. The underlying structure insofar as line-less manufacturing and an object-oriented perspective is regarded remain similar. Figure 2.14 outlines these complex job shop characteristics. Semiconductor manufacturing, due to its technological complexities and requirements, is a prime example for complex job shops (Mönch & Fowler & Mason 2013). However, future trends in globalization will lead to more complex manufacturing settings (Lanza et al. 2019) and, thus, increase the trend towards complex job shops in many more industries.

In complex job shops PPC is likewise more complicated (Waschneck et al. 2016). Hence, a hierarchical PPC decision making approach is typically implemented (Mönch & Fowler & Mason 2013). Nevertheless, on the production control level different lot sizes and routes, stochastic processing times with frequent breakdowns, sequence dependent setup times and reentrant flow (Waschneck et al. 2016) lead to an excessively large solutions space. Therefore, intelligent production control for complex job shops is still in its infacy.

2.2.4 Semiconductor Production Planning and Control

Semiconductor manufacturing decouples PPC decision making in four stages to deal with the complexity (Mönch & Fowler & Dauzère-Pérès, et al. 2011). The decision horizon can be divided into time buckets on all levels (Mönch & Fowler & Mason 2013) as illustrated in

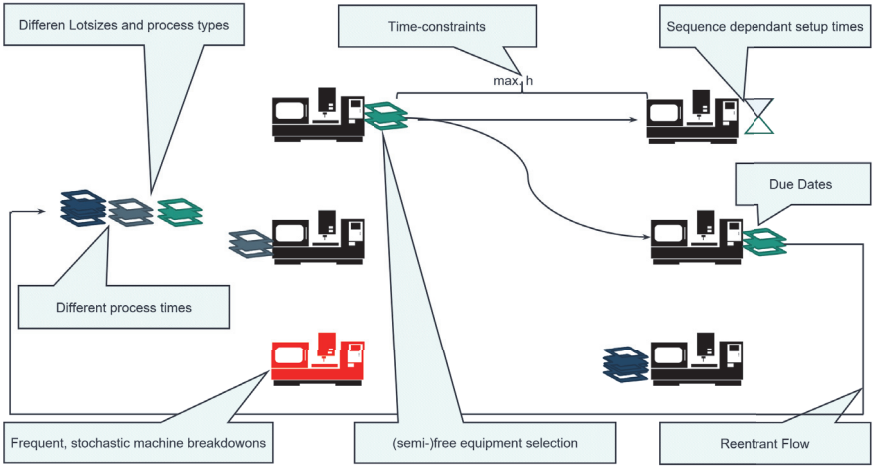

Figure 2.14: Characteristics of complex job shops based on Waschneck et al. (2016) and Mönch & Fowler & Mason (2013)

Figure 2.16. Each time bucket comes with a length and requirement to be achieved by shaping the decisions within this time bucket. In a similar vein to general PPC decision shorter term manufacturing KPIs are estimated based on historical data and used for decision making.

Concretely, planning takes place on a quarterly to yearly level and optimizes for revenue by assuming cycle times and influencing long-term capacity planning and master planning (Mönch & Fowler & Mason 2013). The former plans product mix and associated capacities on an enterprise level in a matter of years. The latter plans production quantities on a more operational monthly to quarterly basis. This leads to production quantities for regarded, typically weekly, time buckets which can be associated to exact semiconductor fabs (Pe 2018). Within these time buckets orders are released, in other words the starting time of production is assigned to each order (Mönch & Fowler & Dauzère-Pérès, et al. 2011). The results are smaller, typically weekly or bi-weekly, time buckets with associated sets of orders to be released into the semiconductor fab. Order releases, intuitively described by Little's Law, directly influence inventory levels, throughput and cycle times (Mönch & Fowler & Mason 2013).

These production planning decisions can be kicked off by numerous triggers (May & Overbeck, et al. 2021). In semiconductor manufacturing event-driven, time-driven or hybrid production planning is prevalent (Mönch & Fowler & Mason 2013). Rolling horizons are the norm in time-

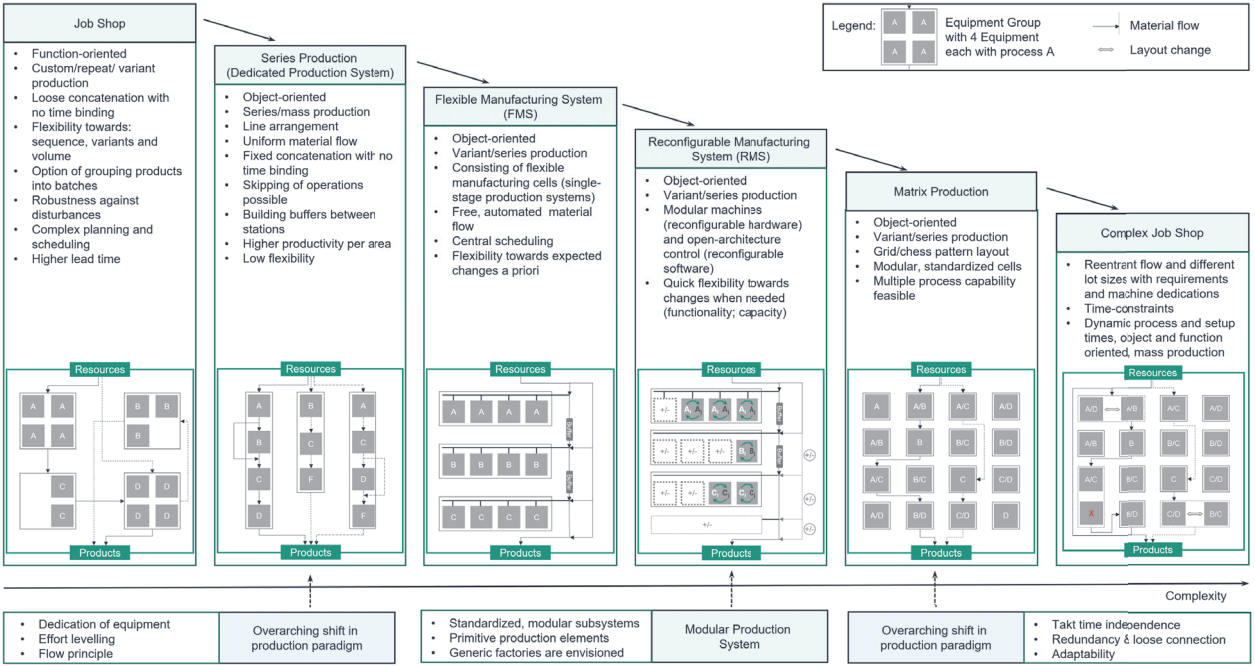

Figure 2.15: Development of complex jobs and complexity comparison with preceding production approaches.

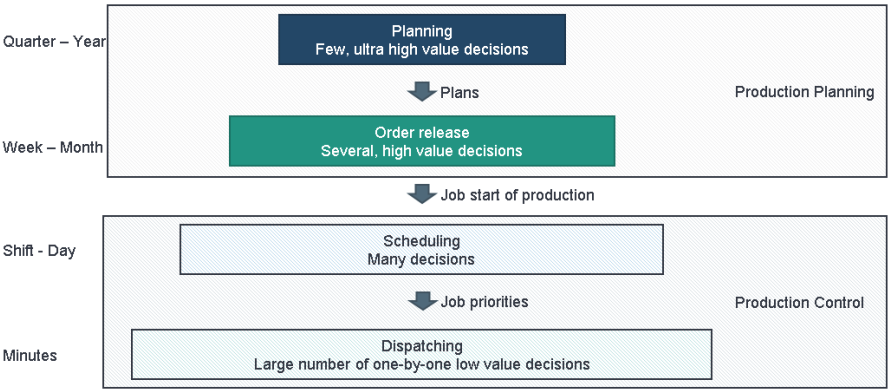

Figure 2.16: Semiconductor PPC hierarchy according to Mönch & Fowler & Mason (2013).

driven production planning. In this case decisions are taken for the planning horizon of several time buckets and in an iterative way adjusted once the first time-bucket is implemented.

Production Control shapes the short term decision within these weekly or shorter buckets. First, based on the known due dates and orders that have to be released, *scheduling* assigns individual jobs to machines (Mason et al. 2007). In general, scheduling allocates scarcely available resources over time to the respective jobs (Mason et al. 2007). Due to the complexity of large semiconductor fabs scheduling is limited to shifts or one day (Mönch & Fowler & Mason 2013). Hence, many decision of medium value have to be taken. Due to the stochastic nature of equipment availability, process quality and timing frequent rescheduling is necessary (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Since semiconductor manufacturing contains reentrant flow, time or space-wise decoupling of these decisions is hardly possible. *Dispatching* uses these desired, scheduled priorities and reacts minute-by-minute to assign available jobs to available resources (Mönch & Fowler & Mason 2013). This includes the possibility to integrate up-to-the-minute information about equipment maintenance, available reticles and further near real-time data. Thus, dispatching has to be performed in a quick and effective way (Wang & Srivathsan, et al. 2018). Priority rules and manual human prescribed priorities are the norm (Mönch & Fowler & Mason 2013). Such dispatching rules are thus typically not intelligent or capable of including sophisticated algorithms or complex problems. The realm of dispatching decisions extends from job to transport allocation, job to machine allocation from buffers and changes in buffer orders, for instance under required maintenance and machine dedications. As to accommodate these quick decisions dispatching processes a large number of typically, individually as low value regarded, decisions (Mönch & Fowler & Mason 2013).

When regarding time-constraints and the high value of scrapped wafers dispatching in fact has a much higher impact and value of decisions (Lima et al. 2021). Due to the current inability of properly incorporating large-scale near real-time data into dispatching necessary to deal with time-constraints human operators are tasked with ensuring time-constraint adherence (May & Maucher, et al. 2021). Hence, future intelligent production control must present solutions to incorporate such near real-time data and intelligent algorithms to address time-constraints to be applicable to semiconductor manufacturing. Solving dispatching under the influence of time-constraints holds great potential. Scheduling restrictions, a typical way of trying to deal with the incapability of dispatching and operators to ensure time-constraint adherence (Klemmt & Mönch 2012), can be eased. Order release can allocate orders more tightly to their due date as time-constraint violations can be better avoided and lengthy restarting of wafer fabrication avoided (Lima et al. 2019). Likewise, by avoiding time-constraint violations less capacity reserve needs to be considered on a pure planning level. Therefore, ensuring time-constraint adherence is paramount for semiconductor manufacturing PPC.

2.2.5 Summary: Solution approach requirements

All in all, production planning and production control within the PPC framework aim at ensuring effective and efficient production on all levels (Mönch & Fowler & Dauzère-Pérès, et al. 2011). Complex job shops, as the climax of complex manufacturing systems, greatly increase the complexity of PPC as explained and visualized in Figure 2.17. While all levels of PPC are affected by the complexity drivers production control is affected most (Waschneck et al. 2016). Designing and implementing a novel production control approach, capable of handling complex job shops, hence, is of major importance. Most notably, time-constraint adherence which is the culmination of complexity drivers should be solved.

To that end, the following requirements must be considered. Firstly, PPC solution approaches must fit into the PPC framework by being able to operate within the prescribed time frames, with the defined performance measures and not compromise on higher level targets as shown in Figure 2.17. Secondly, real-time data should be used to achieve operational targets. Thirdly, historical data should be used to improve decision making and anticipate future behavior of the production system. The underlying rationale is that better informed real-time data based decision, that incorporate behavioral predictions, enable a holistic and effective PPC. Last but not least, PPC in the context of ever increasingly complex manufacturing settings has to be enabled to attain to strict targets without last minute manual, often error prone, human intervention. Time-constraints in semiconductor manufacturing pose such a strict target that cannot be directly controlled with traditional PPC approaches. Therefore, the solutions for complex job shop PPC should contain structural changes compared to nowadays scheduling and priority rules.

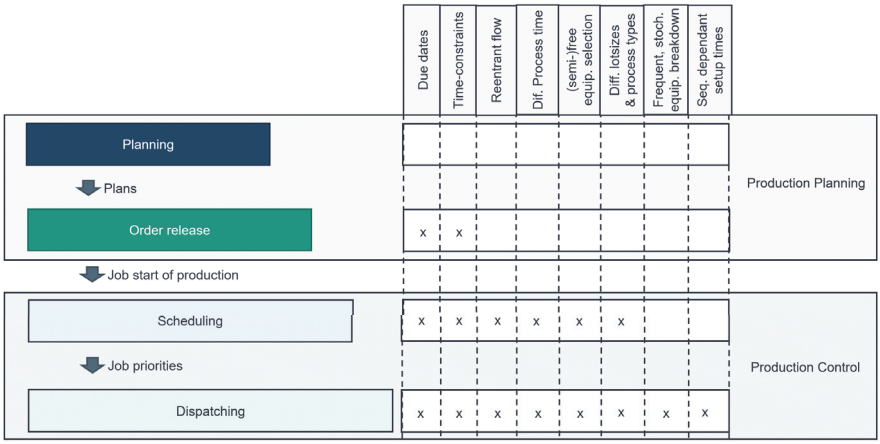

Figure 2.17: Complex job shops strongest PPC complexity drivers.

2.3 Quantitative optimization methods

Optimization takes places under the rationale of object-driven and rational decision making (Nickel et al. 2014). Optimization is closely related to operations research and aims at selecting the best criterion fulfilling element from a set of alternatives (Murty 1994). Relevant criteria include finding maxima, minima or robust values and associated elements (Karsu & Morton 2015). As a quantitative method computational methods are widely applied. In the context of PPC and operations management optimization is often used to find the best objective fulfilling solution under given constraints (Nickel et al. 2014). Alternatively, time series models aim at perfectly modeling a time series through several models (Pham & Kuestenmacher, et al. 2023).

In general, exact methods can be distinguished from approximate methods that find sufficiently good, but not necessarily exactly optimal, solutions (Nickel et al. 2014) as outlined in Figure 2.18. The latter is often implemented in forms of heuristics or meta-heuristics (Burke et al. 2013) which denote fixed decision rules. Heuristics are typically handcrafted to incorporate expert knowledge and patterns observed in systems and hence are suitable even for large problem sizes (Wang & Usher 2004). However, that low required computational effort comes at the cost of typically finding lower quality solutions (Nickel et al. 2014). Instead of handcrafting rules machine learning (ML) as a prominent artificial intelligence methods aims at learning the functional relationship between an input and output based on data (Irani et al. 1993). Therefore, once this relationship is learned it can be applied to smaller and larger

problems with low computational effort and acceptable quality of the solution. Mathematical optimization on the contrary aims at finding the optimal solution of highest quality with typically too high computational effort to be applied to large scale problems (Nickel et al. 2014). The following sections will introduce these methods with a focus on ML as the nowadays most prominent intelligent method.

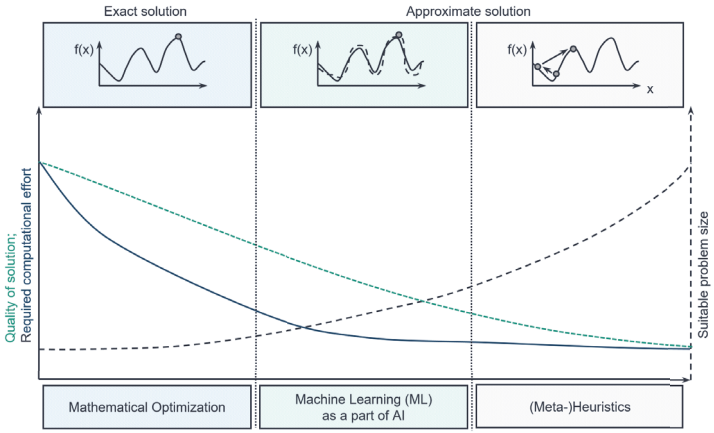

Figure 2.18: Illustrative comparison of exact and approximate quantitative optimization methods (Machine Learning and Heuristics) with respect to suitable size of problems, quality of solution and required computational effort based on Nickel et al. (2014).

2.3.1 Mathematical optimization

Mathematical optimization aims at finding the optimal solution to an optimization problem. Therefore, optimization methods are evaluated, analyzed, improved or developed. The aim lies on improving the evaluation in a given objective function (Nickel et al. 2014). This objective function is to be optimized, in other words minimized (or maximized), and constitutes the first element. Regarding only minimization (or maximization) is sufficient as $f(x_0) \geq f(x) \Leftrightarrow -f(x_0) \leq -f(x)$ holds as $f(x)$ describes the function $f(\cdot)$ over x. Secondly, constraints restrict the solution space. Only solution candidates within the valid space of the decision variable defined in the third step are feasible. (Nickel et al. 2014)

In general, mathematical optimization aims at finding an exact solution with a high quality (Murty 1994). As illustrated in Figure 2.18 this requires high computational effort and is thus only suitable for smaller sizes. The exact complexity, however, depends on the problem type and formulation as introduced in Figure 2.19. The complexity varies between $NP - hard$ for Integer Programming (Murty 1994) and polynomial time for simple linear optimization. The

optimization context decisively influences the exact problem formulation as the parameters that describe the problem can vary between deterministic and stochastic values (Nickel et al. 2014). If the aim is to find robust solutions stochastic optimization is increasingly complex. Constraint Programming regards a different use case in which variables are related through constraints. A selection of the above is presented in the following.

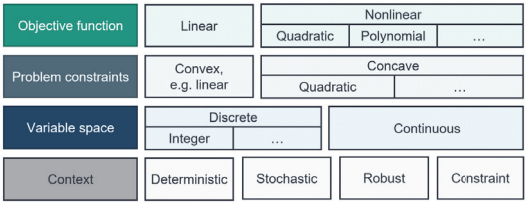

Figure 2.19: Morphological box of a selection of optimization problem types based on Nickel et al. (2014).

Linear Programming regards a linear objective function with convex, linear constraints over a continuous variable space (Murty 1994). The Simplex Algorithm, a simple, efficient algorithm, can solve basic linear optimization problems which include supply chain decisions, production or transport planning (Nickel et al. 2014). In a canonical form a linear problem can be given as:

$$\begin{aligned} max \quad & c^T x \\ s.t. \quad & A^* x \leq b^* \\ & x \geq 0 \end{aligned} \tag{2.2}$$

where x is the decision variable that can be set to maximize the objective function expressed in vector c within the constraints given through the inequalities $A^* x \leq b^*$ and $x \geq 0$.

Integer or mixed-integer programming extends this by restricting the value of decision variables x or a subset of x to integer values (Murty 1994). This vastly increases the problems complexity to $NP - hard$. Real-word problems, however, often fall into this class as resources can often not be split. Examples include sequencing, resource allocation or selection problems (Nickel et al. 2014). In contrast to linear problems which contain continuous linear properties that enable efficient mathematical solution, not present in inequity integer problems (Karsu & Morton 2015). Most common approaches to solve integer or mixed integer problems includes

Branch-and-Bound methods which disaggregate the problem into several branches until optimal integer values are found (Nickel et al. 2014).

Non-linear optimization on the other hand, is complex as the problem is not necessarily convex (Nickel et al. 2014). These problems feature non-linear objective functions or problem constraints which induce possible non corner point optimal solutions. Thus, linear continuous properties do not hold. A feasible approach is the application of a Lagrangian function that transforms constraints into the novel objective function which can be solved with the help of gradient based methods (Nickel et al. 2014).

The setting of these problems can be seen in the context where deterministic models are most widely used. They regard a setting in which parameters, that describe the objective function and problem constraints, in the above example A^*, b^* and c^T, are deterministic and known (Murty 1994). Stochastic or robust optimization however assumes distributions and optimizes over expected values or robustness levels (Nickel et al. 2014). Constrained or dynamic optimization regard optimization from a constraint perspective or recursive, dynamic process perspective. Online optimization extends these different notions to problems where the environment is subject to ongoing information flow (Dunke & Nickel 2016). In nowadays production environments this understanding of finding optimal or near-optimal solutions with limited knowledge about the future is paramount (Dunke & Nickel 2016).

2.3.2 Heuristics and metaheuristics

Heuristics are formulated as a set of rules prescribing how to handle different system states (Kuhnle et al. 2022). In general, heuristics are algorithms that approximate the optimal solution to optimization problems by taking empirical knowledge into account and not mathematically solving the problem (Mönch & Fowler & Mason 2013). Hence, an optimal solution is not guaranteed to be found and the exact solution quality is not known. To provide acceptably good solutions fine tuning heuristics to the exact problem or developing problem specific heuristics is necessary (Burke et al. 2013). The main focus, however, is on finding an acceptably good solution with less required computing power (Waschneck et al. 2016). By storing various rules and retrieving the rules immediate decisions are facilitated and decisions are humanly understandable. Hence, using heuristics are seen as a good compromise between computational speed or effort and quality of the solution for $NP - hard$ problems (Mönch & Fowler & Mason 2013).

Through this high speed decisions can be obtained which are feasible by addressing constraints, such as in integer or mixed-integer optimization, in the set of rules. In general, heuristics can be sorted into procedures to find good, feasible initial solutions to optimization problems and into solutions that improve feasible solutions towards the optimal solution

(Nickel et al. 2014). Heuristics do not search holistically through the solution space, but select among identified feasible solution alternatives. (Burke et al. 2013). Most prominent is the integration of greedy approaches that identify several feasible solutions and greedily select the best alternative until no better alternative solutions are found with the predescribed rules (Yurtsever et al. 2009). This can lead to seemingly shortsighted decision often described as myopic. Heuristics are widely used in industry as they are often formulated by experts, easily interpretable and computationally effective (Waschneck et al. 2016).

Metaheuristics in contrast are algorithms designed to solve problems from a more general perspective (Nickel et al. 2014). To avoid the shortsightedness of heuristics and local optima metaheuristics often introduce randomness or the selection of not only the best greedy solutions. Therefore, general problems, such as scheduling from PPC, can be solved (Nickel et al. 2014). However, the solution found by metaheuristics is not necessarily optimal. Additionally, metaheuristics can hardly recall historical behavior if a defined database is used which decreases the computational speed. Nevertheless, neither heuristics nor metaheuristics can build knowledge and learn from the past (Russell & Norvig 2021).

For short term decision or in the absence of suitable data for the application of these algorithms, humans are often tasked with quantitative decision making in complex production environments such as semiconductor manufacturing (Mönch & Fowler & Mason 2013). Alternatively, artificial intelligence can implement such learning behavior (Russell & Norvig 2021).

2.3.3 Artificial Intelligence

Artificial Intelligence (AI) contains several techniques that draw upon or extend classical mathematical optimization (Russell & Norvig 2021). The underlying approach is to simulate human intelligence based on accurate descriptions. AI aims at artificially creating intelligence demonstrated through learning, reasoning, generalization, knowledge representation and the ability to infer as illustrated in Figure 2.20 (Russell & Norvig 2021). The applications range from language with Natural Language Processing and pictures to high dimensional vectors. Ever since AI's inception in the 1950s and 1960s (Russell & Norvig 2021) the application in production research has been subject to research. Increasing data availability, more potent algorithms and cheaper hardware fuel nowadays plethora of AI applications in practice.

Reasoning regards mainly two tasks. Firstly, reasoning shall use reason on an available body of knowledge to provide answers to questions (Vila 1994). Secondly, this body of knowledge can be extended with new information or checked for inconsistencies (Vila 1994). Doing so, it represents a step-by-step process seemingly copied from human reasoning (Russell & Norvig 2021). It was used to design industrial applications ever since the ability to deal

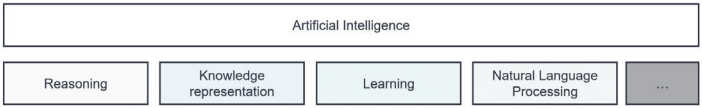

Figure 2.20: Artificial Intelligence and relevant subsets based on Russell & Norvig (2021).

with incomplete information and uncertainty at the end of the 20th century (Vila 1994). The main disadvantages lie in the complexity reasoning over a large body of knowledge and the insufficient knowledge representation. Therefore, modern AI focuses on alternatives.

Knowledge representation aims at representing information intelligently (Russell & Norvig 2021). A Knowledge Base (KB) is used to store this information in form of abstracted knowledge (Studer et al. 1998). In general, the process of designing and managing such KBs has become known as Knowledge Engineering (Studer et al. 1998). Within Knowledge Engineering ontologies have emerged as the most interesting approach (Gruber 1995) as the aim is to create a universally communicable understanding of a domain (Studer et al. 1998). Section 2.3.3.1 introduces ontologies more closely. With the advent of Knowledge Engineering for a semantic web knowledge graphs have become the dominant KB (Jurisica et al. 2004). A knowledge graph has a general structure to store the knowledge within a graph thus allowing flexible relations. A detailed introduction to knowledge graphs is presented in Section 2.3.3.2.

Learning or Machine Learning followed a fundamentally different approach. In contrast to symbolic, knowledge and reasoning based AI Machine Learning as a subsymbolic AI emphasizes statistical relationships (Russell & Norvig 2021). During the AI winter at the end of the 20th century Machine Learning has drifted apart from symbolic AI. Thus, Machine Learning is separately introduced in Section 2.3.4

2.3.3.1 Ontologies

Initially, as a term ontology was used in philosophy to describe the semantic representation of existence (Gruber 1995). While there is no universally accepted definition of an ontology the proposal from Shamsfard & Barforoush (2004) to describe an ontology as a tuple Ont as follows is suitable:

$$Ont = (C, R, A, Top), \tag{2.3}$$

where C represents the set of all concepts, R is the set of all assertions, and A describes the set of all axioms with Top being the hierarchy's highest level of concepts. Concepts refer to notions and the conceptualization of entities. These are represented as the nodes in the ontology network (Shamsfard & Barforoush 2004). Assertions relate any multiple concepts to one another with $H \subset R$ describing the set of taxonomic relations. $B \subset R$ describes non-taxonomic relations which are similarly to H represented as edges in the ontology graph (Shamsfard & Barforoush 2004). To support the understanding and usage of ontology elements axioms about concepts and relations are introduced (Knublauch et al. 2006). The top level hierarchy within the ontology is also mirrored in the relation between ontologies.

Studer et al. (1998) distinguish several types of ontologies. Core ontologies are generic and provide knowledge applicable across domains. As opposed to domain ontologies that contain specific knowledge valid within a particular domain. Combining core ontologies and domain ontologies to solve a problem in a particular area leads to application ontologies. Ontologies can be merged through their concepts and assertions if axioms satisfy certain constraints (Gruber 1995). Thus, reusability, interoperability, flexibility and consistency as well as the possibility to enable reasoning are the strengths of ontologies (Knublauch et al. 2006).

Semantic representations aim at linking semantics to make objects or the web in the semantic web initiative machine readable by building on ontologies (McGuinness & Van Harmelen, et al. 2004). Linked to Extensible Markup Language (XML) the semantic web initiative focused on the Resource Description Framework (RDF) to describe metadata (Knublauch et al. 2006). The extension to include schemata led to the Web Ontology Language (OWL) which serves as a standard ontology exchange format (McGuinness & Van Harmelen, et al. 2004).

2.3.3.2 Knowledge Graphs

While ontologies are well suited to represent knowledge about concepts (Studer et al. 1998) their explanatory power for individual instances is less obvious. Hence, instantiating ontologies increases the ability to explain and model real environments known as a knowledge graph (KG) (May & Kiefer & Kuhnle & Lanza 2022). The advantages include the ability to reason over ontology based KG, the flexibility to the data development in a KG, novel, emerging ML techniques for graphs and the availability of graph query languages that substitute structured query language (SQL) based KBs (Hogan et al. 2021). Therefore, a KG can be defined as a graph containing data to store and convey real world knowledge (Hogan et al. 2021). Herein, entities are represented as nodes and relations between these entities as edges. This paved the way for understanding a KG as a set of triples $(entity, relation, entitiy)$ which corresponds to natural language sentences as a triple of $(subject, predicate, object)$.

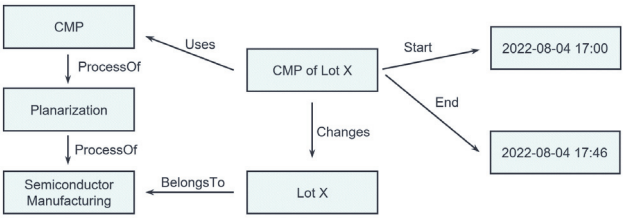

Figure 2.21: Examplary knowledge graph as a direct edge-labeled graph for lot data in a CMP process.

There are multiple graph types used to represent ontologies (Hogan et al. 2021). Based on the RDF schema direct edge-labeled graphs have emerged as the simplest form. Figure 2.21 showcases such a KG, where edges are introduced where the type of the edge is in its label (Hogan et al. 2021).

Big data based knowledge graphs quickly escape from the realm of visually understandable graphs. In order to use that data, graph query languages are used (Pérez et al. 2006). The underlying primitive of those languages are (complex) graph patterns and navigational graph patterns (Hogan et al. 2021). A general graph pattern describes a graph of interconnected constants and variables which are compared against the KG. If the pattern is identified the variables are returned or transformed and returned in complex cases. Navigational patterns can be used to search for paths between two constants within a KG. Coupled with the OWL exchange format that can express KGs (Hogan et al. 2021) query languages such as SPARQL Protocol and RDF Query Language (SPARQL) (Pérez et al. 2006) contribute to the success of knowledge graphs.

2.3.4 Machine Learning

As an alternative to the structured representation of knowledge machine learning aims at the intelligence required to improve in tasks by learning. With the advent of big data machine learning nowadays constitutes a major field of research in both industrial and non-industrial systems (Russell & Norvig 2021). The underlying concepts however date back several decades. In 1957 and 1958 the foundations of modern days most successful machine learning algorithm, neural networks, was laid by Frank Rosenblatt and others who successfully implemented an artificial neural network for the first time in history (Yadav et al. 2015). These early implementations regarded small individual neural networks capable of learning to perform on a narrow set of small problems. In contrast to nowadays specialized, complex and large-scale neural networks, these early implementations of machine learning algorithms suffered under the absence of the huge progress made in computer technology

and the subsequent lacking data and computing power. To handle these large amounts of data deep learning, nowadays dominant approach, applies gigantic implementations of learning algorithms (Russell & Norvig 2021). As a result, machine learning, in other words, machines that learn to perform tasks instead of being programmed to solve tasks has become integral to nowadays manufacturing.

In general, a computer program is considered learning if the associated performance is getting better with increasing experience in that task (Mitchell 2006). The following three classes of machine learning can be distinguished, depending on the data at hand and the objective aimed at:

- *Supervised Learning* learns a mapping function between the input and output of the regarded data that is used to predict the output value for unseen inputs.
- *Unsupervised Learning* detects pattern within a dataset without explicit input and output denomination as for instance widely seen in the detection of anomalies.
- *Reinforcement Learning* places an agent within an environment, based on a Makrov Decision Process, and lets the agent learn a policy to control the environment to reach certain targets.

Beyond the type of ML the exact methods that are used can be distinguished (Russell & Norvig 2021). Therefore, besides introducing supervised learning in Section 2.3.4.1, unsupervised learning in Section 2.3.4.2 and reinforcement learning in Section 2.3.4.3 artificial neural networks and their relevant dependents are explained in Section 2.3.4.4.

2.3.4.1 Supervised Learning

In manufacturing supervised learning is the most commonly applied machine learning technique (Pham & Afify 2005). Based on a dataset containing input data $x_1, x_2, ...x_N$ and the corresponding output data $y_1, y_2, ...y_N$ a model of this association is learned. From this a training dataset $\{(x_k, y_k)$ for $k \in \{1, 2, ...N\}\}$ is selected to abstract the model and consequently apply it to unknown data $x_{unknown}$ and predict the associated output $y_{unknown}$ (Irani et al. 1993). Hence, training is essentially the learning of a mapping function from the input dataspace X to the output dataspace Y according to the inherent structure: $f : X \longrightarrow Y$. An objective function is optimized in an iterative way to improve the performance consistently. If a global optimum is found succesful learning has identified an optimal mapping function with respect to the given dataset. Most notably, the objective function consists of an error function $e : X \times Y \times Y_f \longrightarrow \mathbb{R}$ that describes the cost of falsely associating elements in X and Y and should be minimized. Decision trees, regressional models or neural networks are

most prominently used and often extended through ensemble learning, in other words training multiple models and obtaining predictions through a dedicated voting system. A fundamental approach in all is the association of an input to the class of outputs.

Therefore, widespread applications include classification, i.e. obtaining the expected class associated with the data, and prediction, i.e. estimating the numerical value associated to the input (Russell & Norvig 2021). In manufacturing applications of supervised learning are widespread and include condition monitoring, image recognition, quality prediction for quality control loops or production quantity estimations (Wuest et al. 2016). Supervised learning can also be applied to complement, replace or be an input for heuristics (Irani et al. 1993). The main assumption is that the learning from datasets is a sufficient representation of the underlying data and that unknown data is not sampled from outside the same population (Pham & Afify 2005).

2.3.4.2 Unsupervised Learning

In contrast to supervised learning unsupervised learning lacks the output $\{y_1, y_2, ..., y_N\}$ corresponding to the input. Thus, supervised learning assumes a probability distribution that is ex ante present within the input data $\{x_1, x_2, ..., x_N\}$ and that should be inferred during the learning process as $p_X(x)$. Owing to increasingly large datasets, often labeled as Big Data, unsupervised methods become increasingly important. In manufacturing data is often labeled to facilitate engineering work and, thus, supervised learning is favored (Wuest et al. 2016). However, in wake of increasing data availability through interconnection such as software defined manufacturing (Behrendt et al. 2023) the importance of unsupervised techniques is gaining moment. Most useful techniques detect anomalies, perform clustering or are used in conjunction and preparation of supervised learning (Russell & Norvig 2021).

2.3.4.3 Reinforcement Learning

Instead of datasets Reinforcement Learning (RL) relies on the feedback within an environment to facilitate optimization in decision making within that environment (Koller et al. 2007). RL constitutes a major field in machine learning that is found on the Markov Decision Process (MDP) that models the interaction of an agent within an environment taking decisions to alter the environment as desired. Reinforcing points to underlying learning behavior as knowledge is learned from interacting with the environment and observing the behavior to subsequently adapt the policy used to interact with the environment. Therefore, the learning process is data driven as the reinforcing nature of actions drive the policy towards a better decision making (Russell & Norvig 2021). The underlying concept can be related to that of human learning in the early stage as babies learn through interaction within the environment (Sutton & Barto 2018).

The learner is denoted as an agent that is placed within an environment which follows its own time. While both discrete and continuous time can be selected for the sake of simplicity discrete times as in a regular MDP is assumed for now. Thus, t refers to the current time-step and $t+1$ to the subsequent one. Within this environment the agent perceives the current state $s_t \in S^{RL}$ but does not need to know the previous state s_{t-1} as the markov property holds (Sutton & Barto 2018). Based on s_t the agent selects an action $a_t \in A^{RL}$ it deems helpful in driving the environment towards a desired goal in the next time-step. The next state within this environment is stochastic, depending on the previous state and selected action $P(s_{t+1}|s_t, a_t)$. After having selected the action a_t and depending on the environment's new state a reward r_t is given to the agent. This reward is maximized by the agent during learning to learn an optimal policy $\pi_{RL} : S^{RL} \longrightarrow A^{RL}$ that maps from the current status to the corresponding optimal decision. Through this learning higher and higher rewards are obtained, driving the agent towards the desired behavior. RL can also be applied if the underlying process is not perfectly time discrete or following the markov property (Sutton & Barto 2018). The learning process takes place over time and repeatedly interacting with the environment or in case of closed time horizons over repeatedly interacting through one episode and restarting with a new episode.

2.3.4.4 Neural Networks as universal function approximators

Neural networks form the basis of a plethora of machine learning approaches, in particular used within manufacturing (Wuest et al. 2016). The main advantage is that uncapacitated neural networks can be universal function approximators (Sutton & Barto 2018) so that neural networks and several delineated machine learning algorithms are introduced in the following.

Artificial Neural Network

In a human brain neurons and their interconnectedness are omnipresent which inspired the simplification in the form of Feedforward Neural Networks (FNNs) (Sutton & Barto 2018). FNNs are composed from groups of neurons that are interconnected and only move information forward to the next group. Therefore, no cycles are seen which enables parallelization to improve learning speed. FFNs and typicals architectures are introduced in the following.

Architecture of FNNs

Figure 2.22 visualizes a FNN that consists of three layers, indicating the three different types. First, the input layer constitutes the first layer, which receives the input data x_j. Last, the output layer produces the final neural network output $\hat{y}_i$. In between there are hidden layers which are interconnected and can vary in size and number. The interconnection in a fully

connected FNN is as follows. Each neuron is connected to all neurons in the subsequent layer and each layer has at least one neuron. Imagine two neurons i and j are connected, then in a graph perspective this connection is known as the edge and has the weight factor ω_{ij} denoting the importance of this connection. For calculating the output of a neuron all the input values that are fed forward are used in conjunction with the weight of their edge. Every layer uses an activation function which describes the transformation of the inputs of each neuron. Most commonly rectified linear unit, sigmoid or hyperbolic tangent functions are used (L'Heureux et al. 2017).

As a result, the information that is fed forward and, thus, the information flow in the FNN is controlled through the adjustment of the weights on these edges. Thus, the learning process aims at adjusting the weights in such a way that the error function is minimized.

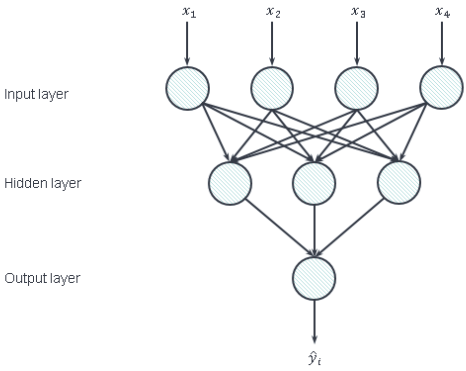

Figure 2.22: Exemplary visualization of the architecture of a fully connected neural network.

Training of FNNs

The iterative training of FNNs takes place in so called epochs. At first the initial weights within the FNN are randomly initialized and improved during the training in the epochs. Within each epoch, the algorithm iterates over the input dataset to train the neural network with the goal to minimize the error, the deviation between the observation y_i and the prediction $\hat{y_i}$ (Russell & Norvig 2021). The derivative for the error function, which is often based on the norm L^2, given the network weights is calculated. Subsequently, the weight adjustment aims at decreasing that error based on a gradient descent approach on the error function surface. This training is subsequently continued over multiple epochs until either the number of maximum training epochs is reached or early stopping is triggered by an increasing error on a holdout dataset. This increase in the validation loss between two training epochs indicates overfitting (Yadav

et al. 2015), in other words memorizing the training data, or a general unstable learning behavior. Besides the error function, the maximum number of epochs and early stopping on validation data, the setting of the learning rate is crucial. It controls the adjustment of the network weights insofar as stepsize of the network weight change can be controlled. To avoid being trapped in a local optimum higher learning rates are used, while lower learning rates stabilize the training process.

If sequential and autocorrelated data is regarded the architecture of the applied neural network can heavily influence its performance. Thus, sequences and autocorrelation should not be simply regarded with standard FNNs that consider observations individually without the preceding time series. To address this challenge, a recurrent element linking the current input and past output was introduced leading to Recurrent Neural Networks.

Recurrent Neural Network

A Recurrent Neural Network (RNN) implements the link between the predecessors of the current input through the realization of cells that store a state in the internal memory. This cell state is changed overtime but in general persistent to capture long term relationships in a long term memory. The major challenge in using RNN lies in vanishing small gradients or too large gradients that tend to explode if working on long sequences (Yamak et al. 2019). To address this challenge two particularly interesting RNNs were developed: Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) neural networks (Cho & Van Merriënboer, et al. 2014). In the following RNNs and their general structure as well as GRU and LSTM are introduced.

Architecture of RNNs

Figure 2.23 visualizes the general architecture of an RNN with x_i being the input sequence, V resembling the loop to the previous output, the used RNN cell and $\hat{y}_i$ resembling the output sequence. Within the RNN cell a hidden state h_i conserves longer term dependency information that only gradually changes over time. The input to the RNN cell is a concatenation or other combination of the loop V and the input sequence x_i. The unfolding of the RNN takes place when the next output value and all its subsequent outputs are not only depending on the cell state h_i but through V also on h_{i-1} and analogously its predecessors. The exact RNN cell has a strong influence on the prediction capabilities of the RNN and in recent times GRU and LSTM have emerged as the most powerful ones (Yamak et al. 2019). Thus, both GRU and LSTM cells are explained and visualized in Figure 2.24 and in the following.

In 1997 Hochreiter and Schmidhuber first introduced LSTM cells (Hochreiter & Schmidhuber 1997) which contains three individual neural networks that act as gates and control the

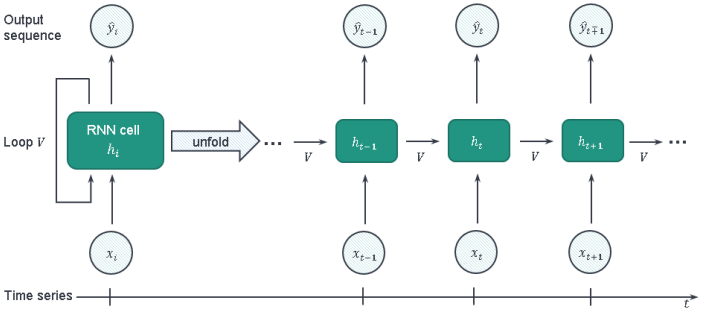

Figure 2.23: Recurrent neural network architecture and unfolding over time based on (Goodfellow et al. 2016).

information flow. As visualized in Figure 2.24 the forget gate erases memories from the cell state c_{t-1}, the input gate adds input from the loop h_{t-1} and input x_t while the output gate controls the final output h_t that also depends on the updated cell state c_t. Finally, the cell state is carried over to the nex period and, hence, serves as the long term memory of the network. This allows persistency of some information, even from the first input x_1 to the last.

In 2014 GRU cells which use a generic RNN input output structure were introduced by Cho & Van Merriënboer, et al. (2014). As visualized in Figure 2.24 LSTMs served as the inspiration for the GRU's internal structure (Yang & Yu, et al. 2020). The decisive difference lies in two details, first the update gate is used to change the persistent state and produce the output. Second, GRU cells do not require the cell state to be transferred to the next periods (c_t). Instead, h_{t-1} the previous hidden state and x_t as the new input value are sufficient to create the so called candidate state which is then used to calculate the next hidden state h_t.

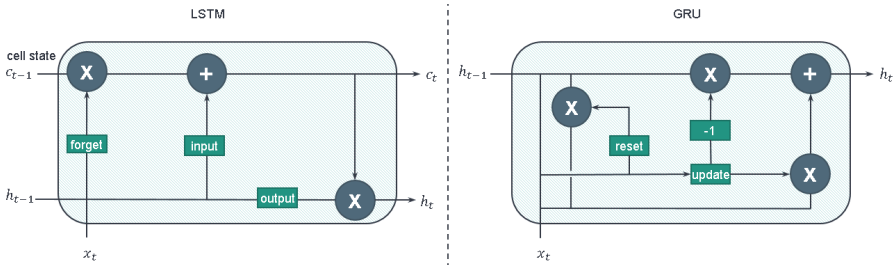

Figure 2.24: Comparison of the cell structure of an LSTM and GRU model.

Training of RNNs

RNN network weight initialization follows the regular FNN training process and applies randomized values. Likewise, the weight adjustment uses gradient descent with backpropagation. The decisive difference occurs in backpropagating the error to the first step. Due to the recurrent nature of RNNs the training can only start from the last state and requires unfolding the RNN to the first input x_0, also known as backpropagation through time, as visualized in Figure 2.23.

Despite their ability to serve as universal function approximators, neural networks and in general machine learning models have limitations that need to be regarded for applications. The following sections hence discusses advantages, opportunities, disadvantages and limitations.

2.3.4.5 Advantages of Machine Learning application

The applicability of machine learning models across domains and application fields is unparalleled (Russell & Norvig 2021). With the availability of more data and freely usable open source software, for instance *Tensorflow*, *PyTorch* or *scikit-learn* the complexities of machine learning applications have eased to a large degree. Thus, in manufacturing machine learning presents a huge chance and exciting opportunity (Wuest et al. 2016). Based on Wuest et al. (2016) and Chen & Sampath, et al. (2023) the potential and crucial aspects of machine learning in the manufacturing domain are put into perspective in the following.

Complex problem modeling ability

Computational problems in manufacturing are often among the most complex problems in the class of $NP-hard$ problems so that ML is a suitable solution technique (Chen & Sampath, et al. 2023) that able to find good solutions in acceptable time for $NP-hard$ problems (Russell & Norvig 2021). As a result the potential of machine learning applications is tremendous and covers all five main KPIs, namely yield, throughput, cycle time, utilization and inventory in direct and indirect ways (Wuest et al. 2016). Since the advent of the third and fourth industrial revolution gathering large data in manufacturing has become the norm (Chen & Sampath, et al. 2023). Therefore, nowadays manufacturing can be seen as rich in data yet poor in the availability of knowledge. ML plays a decisive role in creating this knowledge as it can handle multi-variate, high dimensional data to produce predictions, identify patterns or optimize decision making. Still complex problems require complex models that are more capable in dealing with the highly dimensional data (Wuest et al. 2016). This typically increases the required number of parameters within a machine learning model, moving towards deep learning which increases model inherent uncertainty (Goodfellow et al. 2016).

High adaptability

New data can easily be fed to a machine learning model to enable a quick adaptation which is an important strength of ML (Russell & Norvig 2021). This is promising for complex manufacturing environments, for instance as exhibited in semiconductor manufacturing. Oftentimes changes in the underlying environment can only be identified implicitly through the extraction of the input which can be achieved automatically with machine learning models (Goodfellow et al. 2016). As a result, the required knowledge about the underlying process is less for the application of machine learning techniques than for many traditional approaches (Chen & Sampath, et al. 2023). Additionally, multiple machine learning models can be concatenated or symbiotically used, for instance to first clean and impute data with unsupervised learning and subsequently apply the desired supervised learning.

All in all, ML is a versatile tool with large potential that is comparably easy to apply in the manufacturing domain. Moreover, computational complexity is manageable and requirements on the applied knowledge is minimal.

2.3.4.6 Challenges in applying Machine Learning

Traditional software system challenges are no longer sufficient to describe the challenges and limitations when applying machine learning (Schelter et al. 2015). Due to the increasing application in manufacturing several further challenges arise (Chen & Sampath, et al. 2023). These are introduced in the following section.

Overfitting

The biggest risk of applying machine learning methods lies in the so called overfitting (Domingos 2012). Overfitting refers to cases in which the model learn random features up to memorizing individual data points instead of learning the generalization of the underlying model. Therefore, an ML model that overfits performs very well on the training data set reducing the error to a minimum. As opposed to the application to previously unseen data where results deteriorate (Chen & Sampath, et al. 2023). Domingos (2012) decompose this generalization error into the variance and bias of the model. Wrong assumptions in the model creates a higher bias, in other words moves the overall performance in a structured way, whereas the variance is created by small random vairations in the training data that are incorrectly learned by the model as belonging to the underlying model. As a result, high variance creates a larger variability and reduce the consistency in performance. Both bias and variance are visualized in Figure 2.25. Regularisation is a technique used to prevent overfitting which limits high number of coefficients and punishes too complex structure extraction by the ML model.

Alternatively, boosting or bagging which learns an ensemble of ML models are often applied to reduce overfitting (Chen & Sampath, et al. 2023).

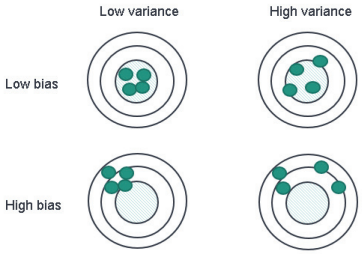

Figure 2.25: Bulls-eye visualization of the trade-off in bias and variance based on (Gudivada et al. 2017).

Curse of dimensionality

The curse of dimensionality states that the machine learning algorithm and hence the model complexity correlates with the dataset dimensionality (L'Heureux et al. 2017). Therefore, applying ML to high dimensional space datasets, in other words using many features as the input and output for each data point, increases the risks associated with the curse of dimensionality which is especially visible with modern deep learning (Goodfellow et al. 2016). The underlying reason is that approximating proper generalization is exponentially more difficult in the dimensionality (Domingos 2012). As ML is essentially using similarities for reasoning, i.e. the ability to relate features to one another and the output, higher dimensional dataset complicate the reasoning. Modern deep learning uses massive computing power to overcome this disadvantage (Goodfellow et al. 2016) or uses attention based mechanisms to focus on important relations (Chen & Sampath, et al. 2023).

Data quality

Data quality is a major obstacle for the application of ML in manufacturing environments (Chen & Sampath, et al. 2023). In general measurement uncertainty, inexact measurements, possibly wrongly obtained input or output data is impossible to avoid due to random errors (Cortes et al. 1994). This aggravated in industrial environments where value creation is focused (Wuest et al. 2016). Thus, obtaining high-quality data for ML application is decisive and should not be compromised. In the following the most common issues based on Zhou & Pan, et al. (2017) and Chen & Sampath, et al. (2023) are listed.

- Data noise refers to missing or incorrect values or outliers within a dataset. Given the typically large amounts of data regarded manually dealing with such noise is infeasible.
- Data redundancy refers to multiple entries of the same underlying data observation within a dataset. The impact can be fatal as the learned model can be skewed. Typical reasons are human errors or inadequate data processing.
- Complex feature representation refers to cases in which the data is not appropriately preprocessed and important features are not selected or made available. While ML can learn features by itself no convergence can be guaranteed so that it is meaningful to use a priori knowledge to craft suitable feature representations.

In a nutshell, the application of machine learning models requires a careful trade-off between the available data, representation and model used. While machine learning engineering is still in its infancy (Chen & Sampath, et al. 2023) knowledgeably interpreting ML training and evaluation is required.

2.3.5 Predictions with Time Series Models

Time series describe a serial data observation over time. In manufacturing environments a big portion of data that is generated comes in the form of time series (Cholette & Djurdjanovic 2014). Studying these time series with time series models is thus of utmost importance (Farahani et al. 2023). Mainly, the description or the prediction of the future time series behavior are regarded. The underlying assumption is that future values in a time series are correlated to its predecessors through an auto-correlation. This chapter introduces time series models, their application to obtain predictions and their extension to a stochastic, interval based understanding in prediction intervals. For instance, once the transition of a product between two consecutive processes on different machines is regarded over a longer period, multiple products take the same route. The result is a time series of transition times. To analyze and understand time-constraint violations from this perspective time series models, associated prediction intervals and possible predictors are introduced in the following and throughout this thesis.

Forecasting and Time Series Models

In general time series X can be regarded as data inherently structured over the realization time of individual values X_t. Using natural numbers as an index, the time series can be written as $X = (X_1, X_2, ...)$. Single and multi-variate models can be distinguished, where uni-variate models base their prediction of the next realization X_t on observations from the same variable X, i.e. $(X_{t-1}, X_{t-2}, ...)$ (Kunst & Wagner 2020). In contrast, multi-variate

models are capable of using multiple variables, typically represented in a vector, to predict X_t. For multi-variate time series models RNNs and machine learning models in general represent good and easily applicable models (Cho & Van Merriënboer, et al. 2014). Thus, the following section focuses on the general introduction of time series and uni-variate models. The most commonly used models Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) (Kunst & Wagner 2020) are introduced in the following.

Autoregressive model (AR)

The AR model represents a random process where the output variable is linear dependent on its prior values. Therefore, the prior observed values (X_{t-1}, X_{t-2}, ...) serve as the input for the linear regression that is used to predict the following value X_t. The hyperparameter p is the order of the autoregressive model and describes the number of prior values used for the prediction. Each prior value is multiplied with the associated parameter $\phi_1, ..., \phi_t$ that is found during the linear regression. Assuming an error term ϵ_t that is not correlated over time with a mean of zero and constant variance the AR(p) model can be given in Equation 2.4.

$$X_t = \sum_{i=1}^{p} [\phi_i X_{t-i}] + \epsilon_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t \tag{2.4}$$

Moving Average model (MA)

In contrast to the AR model the moving average model assumes that the output variable follows a cross-relation to another random variable. It is a stationary random process of order q that evolves around the series' mean μ and has a white noise ϵ_t, in other words a random signal of equal intensities for different frequencies, associated with each observation. By regarding q prior values the model parameters $\theta_1, ..., \theta_q$ are fit with a linear regression according to the given MA(q) model as shown in Equation 2.5.

$$X_t = \mu + \sum_{i=1}^{q} [\theta_i \epsilon_{t-i}] + \epsilon_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p} \tag{2.5}$$

Autoregressive Moving Average (ARMA)

Combining the AR and the AM model yields the weakly stationary ARMA model. Besides extending the applicability to more complex time series Kunst & Wagner (2020) claim that ARMA models are capable of representing time series with less parameters than MA or AR models individually. Thus, instead of building AR and MA models, directly building the ARMA

model is satisfactory and reducing the computational complexity. Inheriting the hyperparameters p, q and parameters $\phi_1, ..., \phi_t$ and $\theta_1, ..., \theta_q$ from the AR and MA model, the ARMA(p, q) model is described in Equation 2.6

$$X_t = \epsilon_t + \sum_{i=1}^{p}[\phi_i X_{t-i}] + \sum_{i=1}^{q}[\theta_i \epsilon_{t-i}] = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_{t-p} \quad 2.6$$

To fit the hyperparameters p and q holistically the Box-Jenkins method makes use of a maximum likelihood estimation (Kunst & Wagner 2020).

Autoregressive Integrated Moving Average (ARIMA)

The generalization of an ARMA model leads to the ARIMA model, a linear, non-stationary and uni-variate random process model. Besides using the two components AR and MA as in an ARMA model, the integration operator I is added to transform the time series to be stationary when differencing. In other words, the differences between two consecutive observations are regarded. In general, the time series can be differenced multiple (d) times and it may include seasonal differenceing where m seasons are eliminated. By doing so the now stationary time series can be solved with an ARMA model. The differenced time series can contain a constant c and is then written as X' and solved according to Equation 2.7.

$$X'_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X'_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \quad 2.7$$

Finding suitable hyperparameters p, d, q is computationally intensive as the additional differencing parameter d greatly increases the search space. Still, the Box-Jenkins approach is a suitable and widely used method to estimate ARIMA models (Kunst & Wagner 2020).

Summary: Prediction Intervals for intelligent production control

Understanding statistics as an interval estimation opens possibilities that go well beyond traditional Production Control approaches that work on discrete or estimated but discrete numbers. When working with forecasts and predictions prediction intervals can estimate the interval in which the next data point is sampled. In manufacturing settings outliers and crucial influences are seen from only one end of the prediction interval, so that a one-sided understanding is necessary. By using a loss function prediction intervals can be optimized and hence effectively used (May & Maucher, et al. 2021). Most prominently, the probability or

the next data point being under or over a relevant cutoff can be used for decision making as visualized in Figure 2.26.

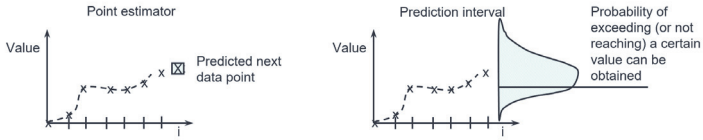

Figure 2.26: Benefit of using prediction intervals over pure point estimation based on May & Maucher, et al. (2021).

Owing to these advantages of using prediction intervals in highly dynamic manufacturing environments and the fact that prediction intervals can be calculated for simple, closed mathematical models such as ARIMA and for ML based neural networks enhances the attractiveness of using prediction intervals in PPC. Thus, in the following the application of prediction intervals should be considered.

2.3.6 Summary: Quantitative methods to optimize production control

Quantitative methods are key to provide and optimize production control as they can handle real-time data and provide decisions beyond human intuition. However, the integration of this data likewise provides a great challenge as traditional mathematical optimization methods are less suitable for handling large scale optimization problems with big data in particular if the problem is of dynamic stochastic nature (Murty 1994) as it is in production control. This shortcoming is widely known and thus heuristics that can easily digest this big data but forego any optimality requirements have become the state-of-the-art (Burke et al. 2013). Additionally, their inability to accrue knowledge and their inflexibility to react to changing environments are the major disadvantages of heuristics. AI can overcome these disadvantages by building up knowledge and increasing the decisional capacity with historic observations. Purely symbolic AI, however, is seen as suffering from a combinatorial explosion for large problem sizes (Russell & Norvig 2021). Nevertheless, the flexibility of storing data and knowledge can be used. Subsymbolic AI and in particular ML has become increasingly popular due to the ability to learn and improve in decision making without suffering from big data. While the training phase of deep learning approaches is computationally expensive their application is comparably fast (Wuest et al. 2016). By extending artificial neural networks to include recurrency and evaluate uncertainties, their applicability has been greatly improved.

When regarding sequential data the presence of auto-correlation, in other words, the correlation of the next observed data point to its predecessors is decisive. In contrast to regular

machine learning approaches that treat each data point individually and equal to one another, recognizing such auto-correlation in time series opens space for the application of time series models. Learning a good fit and representation of these time series is akin to learning in traditional machine learning. Thus, when regarding transition times of lots in semiconductor manufacturing factory, testing for auto-correlation of this sequential data is a must. Sufficient capabilities in modeling these time series can than enable proper production planning and control. Moreover, regarding time series from a machine learning model perspective enriches the number of applicable models.

Therefore, when applying quantitative methods to improve production control with time-constraints the following aspects have to be considered. Firstly, the problem structure has to be analyzed to map suitable quantitative methods to the problem which starts with identification of convex and linear problems. Secondly, the problem size has to be considered in other words the number of variables and constraints that are imposed to the problem has to be manageable. Thirdly, the timely manner of the problem has to be regarded as dynamic problems can be solved with different approaches and different trade-offs between optimality and speed. Last but not least, the relevance of historical information has to be assessed to identify if AI can provide benefits over purely statical, mathematical optimization approaches. In a nutshell, a good indication towards the applicability of individual quantitative methods is required and most suitable ones have to be assessed.

2.4 Production system Digital Twin

A digital twin is in general defined as a virtual, digital replica of the physical original that is interconnected to the physical original (Overbeck & Rose, et al. 2022). Thus, it is closely related to simulations which map a system to a simplified target system replica that is used to analyze the dynamics and create knowledge. As these tools are fundamental to the analysis of dynamic systems their application is widespread in manufacturing. In the following simulations and various extensions of digital twins are introduced from a production system perspective.

2.4.1 Production system simulation

A production system simulation is often used to analyze the dynamic behavior of planned or existing system which is therefore transfered into the digital replica with certain simplifications and assumptions (Ingenieure 1996). There are in general different approaches used that may or may not be combined. Most notably Discrete Event Simulations (DES) with step sizes that are event-based are used (Terkaj & Pedrielli, et al. 2012). Alternatively, system dynamic simulations implement a top down system change approach or agent-based simulations that

use agents to execute decision process are used (Uçar et al. 2020). The latter can consist of a single agent or multi-agent systems. The decentral nature of multi-agent systems makes them advantageous in the application to complex systems but creates the risk of ending in a local optimum (May & Kiefer & Kuhnle & Stricker, et al. 2021).

VDI3633 is a mature guideline on the application of production system simulations that, beyond the modeling type, specifies the following components of a simulation (Ingenieure 1996): External database interfaces to obtain relevant data, a user interface to interact with the human user, a dedicated data management component to deal with the simulation data and finally the simulation kernel that implements the actual simulative behavior. Therefore, the model world is modeled within the simulation kernel and extended with event creation to faciliate the simulation execution (Ingenieure 1996). In combination with production control the interface to external databases and potentially decisional entities is paramount (Terkaj & Urgo 2014).

Due to the capability of DES to model real world production systems accurately, in fact well enough to in situ simulate the behavior with some real time data (Terkaj & Tolio, et al. 2015), DES are often applied for PPC tasks such material flow control or layout planning (Terkaj & Urgo 2015). Therefore, for successful simulations the ability to simply express and model large systems with few assumptions and event is important. DES can additionally be fine tuned to regarded system and scenario as the level of detail is hardly constrained as long as events can be used to model the behavior. Thus, the most common application of DES in production system simulation are the evaluation of ”what if” scenarios (May & Kiefer & Kuhnle & Lanza 2022).

2.4.2 Foresighted Digital Twins

A Digital Twin on a production system level is the convergence of the aim to simulate in situ the production system operations (Terkaj & Tolio, et al. 2015) and the desire to integrate real-time real-world data from digital shadows with system replicas (Uçar et al. 2020). Thus, the digital shadow is the first step toward achieving a digital twin through providing a holistic integration concept for information systems integration in manufacturing (Uçar et al. 2020) and the interface between data storage, warehousing, data collection and their application in a coupled simulation (May & Kuhnle, et al. 2020). The exact implementations vary, however, the overarching principle of creating a two way information flow and control remains (Uçar et al. 2020). By automating the data collection and correlation (May & Kuhnle, et al. 2020) the physical system becomes tangible in the cyber sphere as any point in history of the system becomes apparent. Doing so with low latency creates an up-to-the-minute digital shadow

(May & Overbeck, et al. 2021) that contains the current and previous production system states.

The next logical step lies in using this up-to-the-minute information to analyze the actual system and its behavior in the cyber sphere. This can be seen as a digital twin of the production system where too narrow definition as for instance for product digital twins (Krahe et al. 2022) are avoided. As a definition for this thesis the summary of a digital twin being the digital, virtual instantiation of a physical (unique) asset that exhibits similar properties, behavior and conditions (May & Overbeck, et al. 2021) is used. Therefore, the digital shadow underlies the Digital Twin together with the Digital Master (Liu et al. 2020) that specifies the simulation kernel. As a result, in general many technical implementations are possible as long as several digital twin instantiations can be launched from the digital master and shadow (May & Kiefer & Kuhnle & Lanza 2022). First, this Digital Twin created as an ex post realization is a powerful enabler of analyses of past events. Second, using this highly accurate Digital Twin in anticipation of the near future presents the user with a powerful tool. The latter can be seen as a successor of static simulations and subsequent in situ simulations which can easily be coupled with external data analysis and optimization (May & Overbeck, et al. 2021). The ultimate goal is to use this anticipation within a Digital Twin to create periods of foresight into the short- and medium term behavior of the production system that can then be exploited by the production control (May & Overbeck, et al. 2021). Thus, foresighted digital twins must implement more realistic system environments in the simulation core than conventional simulation software can do.

2.4.3 Knowledge Graph based Digital Twins

Based on the need of digital twins to enable a real-time behavior anticipation (Overbeck & Rose, et al. 2022) and their requirement to provide interlinkage with the real world production system (Liu et al. 2020) the need for up-to-the-minute digital twins was born (May & Overbeck, et al. 2021). In light of enabling foresighted digital twins traditional, static DES are not longer suitable. To address this issue integrating production system data-sources to an ontology and knowledge graph that is integrated into the production system simulation to create the digital twin was proposed (Calvo et al. 2023).

Different Ontology Integration Levels for production system simulations and digital twins were proposed (May & Kiefer & Kuhnle & Lanza 2022):

Level 1: Information about the structure and capacity, e.g. machines, material flow and queues, is contained in the ontology.

Level 2: Process specific information, such as process times, setup or maintenance times and behavior, is included in the ontology.

Level 3: Products and their respective production plans are included in the ontology and hence enable the core required features for a simulation.

Level 4: By including historic and current event information in the ontology the digital shadow is fully implemented and a simulation model filled.

Level 5: All information to implement an up-to-the-minute digital twin, namely all of the above and the general simulation specification, is contained within the ontology or knowledge graph.

The first levels of Ontology Integration Level have been used in the past to create virtual factories for in situ simulations (Terkaj & Tolio, et al. 2015) and foresighted digital twins (May & Kiefer & Kuhnle & Lanza 2022), to create a manufacturing domain ontology (Mönch & Stehli 2003; Mazzola et al. 2016) or to represent a semiconductor fab (Schulz et al. 2022). By reaching Ontology Integration Level 5 a full digital twin based on the real counterpart with up-to-the-minute behavior is implemented (May & Kiefer & Kuhnle & Lanza 2022). By reaching Level 5 a foresighted digital twin is achieved as with small variations the knowledge graph based digital twin can be run multiple times to analyze production planning and control decisions. An exemplary visualization of the interconnection between the real, physical system and the knowledge graph (KG) that underlies the simulation is presented in Figure 2.27. The real system and its entities is directly reflected with a knowledge graph. The structure of the KG is based on an ontology that can describe the object of study, here a production system, accurately. However, due to the flexible, ontological approach, the KG and ontology in the core of the digital twin and simulation can technically be selected freely as long as the simulation framework is adapted or the basic structure modeled in the ontology is adhered to (May & Kiefer & Kuhnle & Lanza 2022).

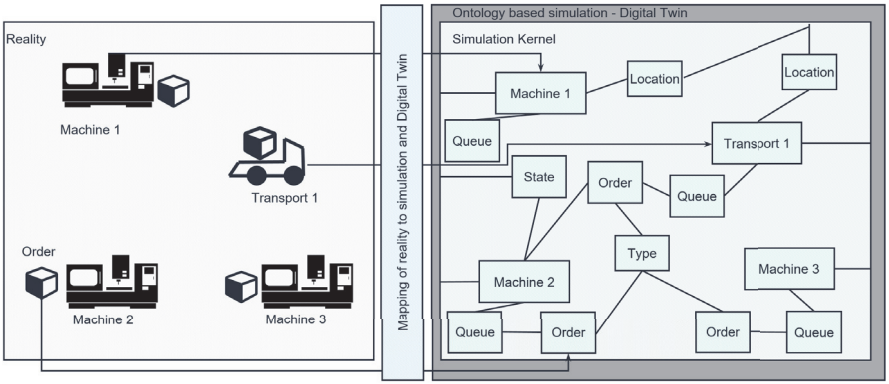

Figure 2.27: Interconnection of an ontology and knowledge graph based digital twin with its physical counterpart based on May & Kiefer & Kuhnle & Lanza (2022).

By basing the simulation within the digital twin on a KG with a manufacturing domain specific ontology several advantages evolve (May & Kiefer & Kuhnle & Lanza 2022). Firstly, the state of the real system digital twin or simulation can be used interchangeably. This enables inference towards the real world use-case and allows storage and configuration during optimization. Secondly, this KG is of highly dynamic structure so that it can be extended or in general changed during the simulation or digital twin runtime. This is important to incorporate real-world changes, such as breakdowns or layout changes, without the need to handcraft the system or state. Thirdly, the simulation core and the behavior can be generalized to the needs of different settings and production environments, e.g. for batch processing. Most notably, product inherent constraints such as time constraints can be included within the KG and with a simple SPARQL query can be verified. This leads directly to the fifth advantage of enabling a strict decomposition of simulation framework and behavioral production control that can interact with the simulation framework in a service oriented manner. The latter can be completely decoupled as the ontology and its current status and data in the KG can be queried instead of relying on programmatically predetermined values that are available in traditional DES environments. Last but not least, by saving the simulation state in the KG and likewise saving future and past events the simulation becomes iterable. In other words, one can go back and forth stepwise and at the same time make changes to the system representation. This enables shortcuts in branch and bound or tree based optimization approaches as dominated solutions and unattractive paths can be cut off. (May & Kiefer & Kuhnle & Lanza 2022)

The availability of open source simulation software that is based on such an ontology, knowledge graph approach in form of OntologySim made available by the author of this thesis (May & Kiefer & Kuhnle & Lanza 2022), enables extendability. Most prominently, the adaptation to a use-case can be seen as much faster and much more stable than implementations from scratch. As KPIs are integrated based on a standardized approach behavior can be compared comprehensively.

2.4.4 Summary: Simulations as Digital Twins for production control

In PPC dynamic systems are regarded which can hardly be understood with purely static, average based methods. Therefore, DES as a simulation tool has gained momentum and has been widely applied. By tracking event and introducing changes to the system of study or its environment through time discrete events the system's behavior can be examined. Thus, DES are used regularly for production planning and selecting production control in longer term perspectives (Terkaj & Tolio, et al. 2015). Extending this simulation through an interlinkage with data from the real system and its digital shadow, that stores this real-time data, creates the digital twin (Overbeck & Rose, et al. 2022). The advantage of a real-system interlinked digital twin is that the actual behavior with the real jobs, breakdowns etc. can be studied and a high fidelity understanding can be developed. Transferring the traditional approach of evaluating planning and control measures to short-term decisions in a digital twin can be achieved with a foresighted digital twin (May & Overbeck, et al. 2021). For the application in complex manufacturing systems with high degrees of flexibility the implementation based on a KG and ontology, such as in OntologySim (May & Kiefer & Kuhnle & Lanza 2022), provides additional benefits as complex systems and their behavior can be flexibly modeled and connected to the real system. Therefore, throughout these developments, ever shorter decision intervals become possible, ever increasing accuracy allows better result predictions and, hence, the benefit of transitioning towards these digital twins increases. Figure 2.28 illustrates this relationship. Therefore, the main advantage is that the knowledge graph based digital twin can be used to improve, select and validated production control performance on short notice in a flexible way. Hence, now the effects of decisions on complex systems can be studied.

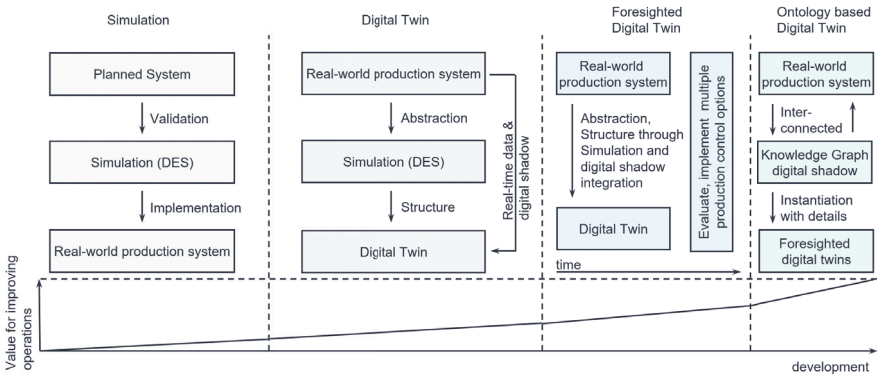

Figure 2.28: Evolution and increased benefit from simulation to foresighted and ultimately ontology-based Digital Twin for a production system.

3 State-of-the-art literature review

Intelligently controlling complex manufacturing systems has been studied for years (Paternina-Arboleda & Das 2001). Various approaches, ranging from reducing complexity in the system through lean management, solving large scale optimization problems to dynamic, intelligent decision-making, have been proposed. Acknowledging conflicting targets and integrating dynamic restrictions becomes necessary in complex job shops to effectively control the production. Machine Learning as means to leverage on both optimization and the ability to improve over time holds great potential. Likewise, knowledge informed decisions and integrative decision making has to be regarded. For intelligently controlling complex job shops with time-constraints, the respective state-of-the-art research is introduced in the following Section 3.1. Section 3.1.1 presents digital twin based approaches for intelligent production control with a focus on the integration of static and dynamic knowledge integration as necessary for complex job shops. Intelligent production control to handle time-constraints on a planning and scheduling level is introduced in Section 3.1.2. Dispatching under time-constraints as short term intelligent production control is presented in Section 3.1.3. Section 3.1.4 introduces learning based production control with time-constraints. Concluding this literature review Section 3.2 presents the research deficit that is addressed within this work.

3.1 Literature review of focus areas

Research conducted in fields related to this work, as shown in the following paragraphs, can be analyzed and classified according to different dimensions. The selection of this research is based on a rigor literature review process based on the grounded theory approach by Wolfswinkel et al. (2013). Focusing on these dimensions the research deficit can be analyzed.

System scope: Complex job shops are characterized by high volume high variant production and exhibit the properties outlined in Section 2.2.3. As the complexity is magnified with increasing system size and the exact degree of complexities regarded in a system driven by stochastic processes, the setting and regarded system size has to be regarded. A major industry that exhibits complex job shop properties and that suffers significantly under time-constraints is semiconductor manufacturing. To properly address the identified research questions the search space is restricted to a semiconductor manufacturing scope with time-constraints.

Modeling: As the system regarded has to be abstracted to be understood and controlled, the exact modeling approach is crucial. The approach developed in this thesis needs to be

adaptive, executable in near real-time and connectable to real-world data to not delay production. Within the realm of data-based modeling, explicit modeling approaches, experimental studies and analytical or meta models can be used. The modeling choice heavily influences the solution technique that can be used.

Solution: Based on the modeling approach, the solution technique plays a pivotal role in performance and applicability to the selected problem scope. Exact and approximate solutions can be distinguished. Their requirements on consistency, extensiveness and completeness about data and the assumptions used during modeling and solution process vary. Thus, the confidence in the results and the required computing effort vary alike.

Objective: Production control has to comprehensively evaluate different objectives. The focus and intensity of regarding specific objectives, however, can vary vastly. In order to play a significant role in the research within this thesis, time-constraints and general targets such as the KPIs introduced before within the complex job shops shall be regarded. A dynamic evaluation should be evaluated and, ideally, time-constraint violations minimized. The result of this present work should follow a research approach in which the validation suggests the transferability to other complex job shops and production control problems.

3.1.1 Digital twins for intelligent production control

Digital twins that incorporate near real-time data from the real system and digital shadow into a simulation model on production system level can improve production control (Uhlemann et al. 2017). Their benefits range from evaluating novel production control approaches in the digital twin (Negri et al. 2017), to creating data in a digital twin that can be used in production control (Uhlemann et al. 2017) up to coupling production control within a digital twin and improving and selecting future production control (May & Overbeck, et al. 2021).

Uhlemann et al. (2017) describe a general approach of utilizing digital twins on a production system level. By focusing on planning approaches, the authors praise the benefit of being able to transform manual, intuition based processes to data based, dynamic processes, in a similar, but more accurate, vein to DES. This allows handling of complexity as data interconnectedness reduces manual effort for handling complexities. The approach is transferred to production control to validate performances and obtain data that is hardly accessible and interpretable from direct shopfloor sources. The approach is focused on SME with medium complex production systems.

Negri et al. (2017) propose the entanglement of simulations on different levels through the digital twin. Doing so can evaluate the performance of different production control approaches in real-time. They envision synchronized simulations in a digital twin to overcome

the reality-to-simulation gap. This gap hinders effective implementability and assessment of production control between the virtual and real world. The approach focuses on a wide range of production systems but lacks a concrete application.

Min et al. (2019) implement a digital twin of a petrochemical factory and use real-time digital twin data to start a pipeline of machine learning approaches that aim to be kept up to date. The foremost goal is to synchronize chemical production to minimize waste and waiting. They show that digital twins exhibit benefits when dealing with high dimensional data, time lags and synchronizing time series data. The effectiveness is examined and the authors foresee similar application in other industries.

May & Kuhnle, et al. (2020) research the applicability of digital twins, digital shadows and related concepts in the PPC framework. The highest benefits comes from coupling operational PPC with a digital twin and relaying the analyzed behavior to tactical decisions. They regard a complex job shop in semiconductor manufacturing with a handful of machines aiming to improve dispatching. Within the digital twin, different heuristics and reinforcement learning are compared. Based on this comparison and additionally available data, novel, better heuristics and reinforcement learning agents are delineated and partial autonomy is implemented.

The concept of implementing reinforcement learning for production control decisions such as material flow (Overbeck & Hugues, et al. 2021) or maintenance (Hoffmann et al. 2021) embedded within a digital twin is frequently regarded. Kuhnle et al. (2022) use a digital twin of an area within a semiconductor manufacturing system to train reinforcement learning and regard the explainability of decisions to improve both the reinforcement learning and traditional production control. Doing so, they aim to overcome the limited understanding of black box machine learning models which creates distrust in such systems. Valet et al. (2022) model a real-world semiconductor manufacturing plant in a digital twin and optimize for opportunistic maintenance decisions with deep reinforcement learning. The digital twin greatly improves production control performance in this study.

May & Overbeck, et al. (2021) envision the real-time evaluation of production control in the near-future within a foresighted digital twin. They realize the digital twin coupling with a simple matrix production system and improve production control performance by selecting specialized heuristics or reinforcement learning for short time periods depending on the up-to-the-minute suitability instead on averaging over longer periods. A generalized version, OntologySim, with knowledge graph based data integration is presented by May & Kiefer & Kuhnle & Lanza (2022). The integration of real-system alike data, extensive interconnection, a wide range of standardized KPIs and free foresight triggering benefits production control.

In a nutshell, the virtual system replica in form of a digital twin that serves as a playground to augment real world data, evaluate and improve production control is pivotal. Nevertheless, the application in semiconductor manufacturing is far from real-world applicability. Additionally, the integration of all complex job shop's complexities, in particular time-constraints, is lacking behind and no studies that do so have been identified. As a result an in-depth discussion and tabular analysis is skipped. The general approach of digital twin is interesting to integrate real-time data and glimpse into the future. Therefore, using digital twins to improve time-constraint adherence is promising, despite not having been studied so far.

3.1.2 Dealing with time-constraints in capacity planning and scheduling

Following the PPC hierarchy, this section introduces the state-of-the-art in dealing with time-constraints on a production planning level in capacity planning in Section 3.1.2.1. In a similar vein, on the production control level, implementing tactical and operational PPC, scheduling levers on ensuring time-constraint adherence are regarded in Section 3.1.2.2.

3.1.2.1 Capacity planning

The following presents an overview of the state-of-the-art in capacity planning approaches that aim at or incorporate time-constraint adherence. Regarded studies are directly or indirectly reducing time-constraint violation. The results are presented and clustered according to the modeling and solution approach.

Queuing systems for capacity planning to reduce time constraint violation

Using queuing systems to model a complex job shop and its manifestation, i.e. a wafer fab, is common. Robinson & Giglio (1999) predict the time constraint violation probability derived using queuing theory. Following Kendall's notation, a system of c Machines, denoted as $M/M/c$, has a memoryless, Poisson arrival process and similarly a memoryless, exponential service time. The authors regard a two element system. In a similar vein, Tu & Chen (2009a), Tu & Chen (2009b), Tu & Chen (2010), Tu & Chen (2011) and Tu & Liou (2006) use a $G/G/c$ queuing model with generally independent service time and arrival times to develop a model to determine the optimal capacity. They investigate the resulting queuing network to identify the required number of machines under prescribed yield targets. Machine breakdowns are included by Tu & Chen (2009b), who calculate the probability of wafers violating the time-constraint. The approach can be extended to include batch equipment (Tu & Chen 2011). Ono et al. (2006) assess tool group's time-constraint violation risk with queuing models to improve capacity at tool groups which are critical. As opposed to Kitamura et al. (2006) who restrict work in progress (WIP) inventory between operations to derive an optimal capacity plan that minimizes time-constraint violation. Decisive lower level production control, convoluted

material flow, breakdowns, clustering tools or batching processes and machine dedications are not regarded.

Experimental evaluation of time-constraint influence on capacity planning

Time-constraint influence on throughput and yield is studied by Pappert et al. (2016) who simulate a complex job shop accounting for time-constraints but not controlling time-constraints. Time-constraints led to a large loss in throughput and yield and, hence, require an increased capacity. In contrast to the technological requirements introduced in Section 2.1 they advocate for reducing time-constraints or increasing the time limits as much as possible. In a previous study Huang et al. (2011) quantified the effects of time-constraints with similar conclusions. Therefore, relaxing these constraints and properly dealing with time-constraints on a production control level is key to reduced capacity and increase quality.

Machine Learning influence on capacity planning under time-constraints

With increasingly available real-time data from wafer fabs, the application of ML becomes more beneficial. On a capacity level, Kuo et al. (2011) use an artificial neural network to predict the WIP inventory levels on a tool group level. Through a sensitivity analysis the influence of technical improvement measures, e.g. increased tool availability to reduce time-constraint violation, on WIP inventory levels or cycle time are quantified. With real-world data from a Taiwanese wafer fab the method for cycle time reduction is validated. Extending this work with the effects of tool allocations, Chien & Kuo, et al. (2020) prove the influence of capacity planning decisions on time-constraint adherence.

Capacity planning summary

All in all, integrating time-constraints into capacity planning considerations is paramount to create suitable conditions and prevent poorly designed systems and capacities. The studies are nevertheless not applicable to ensure time-constraint adherence throughout as longterm probability distributions, approximate WIP inventory levels and assumptions on production control are implemented. The models hence work very well to balance a trade-off between time-constraints, costs and other operational KPIs as they on average allow good estimations. However, individual time-constraint adherence cannot be regarded. Therefore, operational production control is required to ensure time-constraint adherence on an operational level (Ono et al. 2006).

3.1.2.2 Scheduling

Scheduling is an operational production control approach that allocates jobs to machines under certain constraints. The typical time span of scheduling in complex jobs ranges up

to hours or days but does not regard real-time decisions. Due to that abstract, prescriptive nature it desirable to first regard scheduling for production control in complex job shops and semiconductor manufacturing (Mönch & Fowler & Mason 2013). To achieve feasible schedules in a dynamic, uncertain and complex manufacturing environment, the ability to find solutions of acceptable quality in short time is pivotal. Dynamic rescheduling has to be performed frequently so that waiting for long solution times is infeasible. In the absence of sufficient data and powerful algorithms Bixby et al. (2006) formulated the hypothesis that scheduling was more desirable to dispatching to control time-constraints in complex job shops as an analytical approach is used to achieve coordination and control. Therefore, the majority of research regarding time-constraints is on a scheduling level. Due to the high complexity scheduling has not solved time-constraint adherence in complex job shops but it is still under ongoing research. This research can best be clustered based on the modeling and solution approaches used. The regarded problem size is paramount as large scheduling problems are hardly solvable in acceptable time.

Exact solution approaches for scheduling with time-constraints

As lots in complex job shops can hardly be split freely, scheduling regards them as binary allocations which leads to Mixed Integer Linear Programs (MILP) as efficient, linear modeling approaches. Chen & Yang (2006) solve MILP open shop, job shop and flow shop models for optimality with few machines. They aim to reduce makespan under time-constraints. Similarly, Klemmt & Horn, et al. (2008) reduce cycle times and aim at minimizing waiting times. Kao et al. (2011) reveal in a comparable study 18.5% utilization increase without sacrificing time-constraint performance. However, these approaches regard only partial complex job shops as they are restricted to furnace tools to preserve manageable computing effort. This restriction in size is mirrored in the number of jobs that can be scheduled (Yu & Kim & Jung, et al. 2013). In a relaxed Mixed Integer Programming (MIP) approach Yu & Kim & Jung, et al. (2013) minimize waiting time variation to increase time-constraint adherence for up to 25 jobs. Bigger problems are addressed with approximate solutions. An et al. (2016) can schedule up to 30 lots in two machines with time-constraints by applying heuristics to find initial solutions improving bound computation efficiency in a branch-and-bound algorithm. A similar technique is used by Kim & Lee (2017) who can schedule 20 lots over three machines.

Beyond traditional time-based models Cho & Park, et al. (2014) compare a slot-based formulation where the allocation is based on slot positions. They identify the need to consider both times and slots as the results depend heavily on the problem parameters. Another change, proposed by Maleck & Eckert (2017), successfully applies constraint programming to scheduling under time-constraints. In a real, small scale production setting constraint programming found solutions faster and resulted in better overall performance. They build

upon the previous study by Maleck & Weigert, et al. (2017) who considered machine break downs and reschedule hourly or during breakdowns and interruptions. Maleck & Nieke & Bock & Pabst & Stehli (2018) refine the approach and apply it to a larger area in a complex job shop.

Decomposition based scheduling with time-constraints

In contrast to branch-and-bound, decomposition methods divide the large scheduling problem into smaller, individually solved subproblems. Through composition the overall schedule can be obtained. A three level decomposition is proposed by Sun et al. (2005) to minimize time-constraint violations and by Maleck & Nieke & Bock & Pabst & Schulze, et al. (2019) to minimize cycle time. Bixby et al. (2006) use a recursive space-time allocation and deploy their multi-objective algorithm in a real wafer-fab. Total tardiness is addressed by Klemmt & Mönch (2012) with recursive near-optimal solution. Jung et al. (2013) and Jung et al. (2014) use a similar approach to reduce time constraint violation rates and improve cycle time by focusing on diffusion processes. Decomposition is performed along furnace tools and solved with branch-and-bound.

Meta-heuristics for scheduling with time-constraints

To overcome the computational effort disadvantage of exact solution approaches meta-heuristics often combine construction heuristics to find initial good solutions with neighborhood search to improve these solutions. To address time-constraints in control of complex job shops Su (2003) were the first to use a two-stage simulated annealing based algorithm to minimize makespan. First, lots are allocated to batches and secondly the sequence within these batches is determined through an interchange mechanism. They regard time-constrained transfer from a batching equipment to a single tool. Yugma et al. (2012) propose a disjunctive graph representation and solve this batching and scheduling problem through job insertion, iterative sampling and simulated annealing. The object of study is limited to the diffusion work area in a wafer fab. In the same setting Zhou & Wu (2017) apply a greedy construction heuristic and simulated annealing. Similarly and in a recursive manner Nattaf et al. (2019) apply heuristics and simulative annealing to both MILP and constraint programming models. For larger problems, yet not regarding whole semiconductor fabs, Zhou & Lin, et al. (2019) propose a cuckoo search algorithm as an evolutionary strategy that outperformed previous approaches.

Genetic algorithms are alternative evolutionary meta-heuristics that have attracted many researchers in controlling time-constraints. Mason et al. (2007) regard two time-constrained batch processing steps and solve the multi-objective problem with NSGA-II and a genetic algorithm with different batching rules. In the furnace work area Chien & Chen (2007) solve

batch sequencing with genetic algorithms. Similarly, Klemmt & Horn, et al. (2008) apply genetic algorithms to oven batching problems in a semiconductor fab. Later, Jia et al. (2013) use a genetic algorithm for closed loop scheduling of parallel batch machines. Time-constraints are controlled best by a subpopulation based genetic algorithm in a study by Wang & Chien, et al. (2014). An improved genetic algorithm from Wang & Chien, et al. (2015) learns the probability distribution in the population to outperform alternative heuristics and meta-heuristics. Most recently, Lee (2020) harness increased computational power to grid search parametrization of a genetic algorithm that outperformed simple heuristics and simulated annealing in reducing total tardiness and hence minimizing time-constraint violations as a secondary target.

Heuristic scheduling under time-constraints

Faster decisions can be obtained from pure heuristic approaches that rely less on analytical optimization but can be crafted based on simulative results. While many heuristics implicitly regard time-constraint through WIP inventory control, they are rarely directly regarded within heuristics as the complexity of these heuristics increases dramatically. Opposed to their simulative annealing based algorithm Su (2003) additionally used simple heuristic. Only built on heuristics Ham & Raiford, et al. (2006) create a rule based wet etch to furnace transitioning algorithm. Based on a priority list from the furnace the wet etch scheduling serves both the furnace and other work areas. The most successful real-world implementation is presented by Yurtsever et al. (2009) who regard batch scheduling in the diffusion work area. They regard multi-objective optimization under constraints, in particular time-constraints. The heuristic focused on minimizing idle time between batches through priority based batch selection. Results in a wafer fab in Austin, Texas, US decreased cycle times by up to 25% and increased throughput up to 10%. Nevertheless, time-constraint adherence performance is not reported. As opposed to Yu & Kim & Lee (2017) who regard a generalized two-machine flowshop aiming to minimize time-constraint violation through waiting time variation control.

Other modeling approaches that use heuristics are shown by Perraudat et al. (2019) who build a decision support system based on a kanban system model of a regarded subsystem. Wu & Zhao, et al. (2016) however use an analytical model. They aim at quantifying the trade-off between high capacity and low time-constraint violation and rework to control WIP inventory levels.

Scheduling summary

All in all, time-constraints are frequently integrated into scheduling as direct or indirect objectives. Scheduling without considering time-constraints is leading to large scrap rates (Maleck & Nieke & Bock & Pabst & Stehli 2018) or reduced utilization and throughput if inventory levels are artificially held low (Kao et al. 2011). Besides Yurtsever et al. (2009) the scheduling

is restricted to few machines or one or two work areas, larger complex job shops as they occur in semiconductor manufacturing fabs are not solved. The main disadvantage of the proposed scheduling algorithms is that they seldom explicitly aim at ensuring time-constraint adherence without sacrificing operational performance. Due to the highly dynamic nature of complex job shops, prescribed schedules are quickly impossible to implement (Maleck & Weigert, et al. 2017). Heuristics however are hardly capable to control these complex systems or time-constraint violations are accepted for operational performance as in Yurtsever et al. (2009). Thus, to effectively deal with intelligent production control under time-constraints, dispatching decisions have to be controlled in real-time. However, current heuristics, while fast in decision making with acceptable operational results, are incapable of effectively including the complexities of a complex job shop into the decision making process.

3.1.3 Adhering to time-constraints in dispatching in complex job shops

To integrate real-time information and decide quickly, dispatching as the lowest PPC level allocates jobs to available machines. In the context of complex job shops and semiconductor manufacturing lots of wafers are assigned to available resources for processing. Therefore, dispatching includes a gate-keeping decision for each lot (May & Behnen, et al. 2021) as this lot does or does not continue to wait for processing or transport. Due to the dynamic and hardly predictable nature of complex job shops, the schedule provided from scheduling is implemented or changed during the dispatching. Therefore, quick reactions are required and up-to-the-minute information about availability, breakdowns etc. has to be considered. To address this speed, priority rules are widely used (Altenmüller et al. 2020) despite their inability to regard global optimization perspectives (Valet et al. 2022). The available body of knowledge can be clustered according to the modeling and solution approach.

Simulation based dispatching control

Reducing the impact of time-constraints is studied by Scholl & Domaschke (2000) who use a simulation to research time-constraint impact between wet-etch and furnace operations. They identify impact reduction measures to ensure less time-constraint violations through relaxing of the lot order system depending on the time-constraints. Similarly, Pappert et al. (2016) regard time-constraint influence in a coating work are with simulations to capture the dynamic nature. Based on this investigation Zhang et al. (2016) introduce a dynamic control policy heuristic to balance time-constraint violations and cycle time. Through this dynamic handling average cycle times remain constant whereas time-constraint violations can be reduced. Lee & Chen, et al. (2005) successfully introduce a five step control chain based on WIP inventory levels and tool availability to easen the pain of time-constraints in a complex job shop. Another rule based dispatching policy validated in simulations is provided by Tu & Chen & Liu (2010)

who dynamically control WIP inventory and indirectly time-constraint adherence. Toyoshima et al. (2013) further consider product-mix as a relevant factor. As opposed to Kobayashi et al. (2013) who speed up time-constrained lots to increase adherence and improve throughput in a reentrant job shop. Arima et al. (2015) follow up on this study through a combination of loading rule and dispatcher. Processes in front of individual time-constraints are regarded by Ciccullo et al. (2014) and Pirovano et al. (2020) who consider the gate keeping decision and batching in a cleaning process with time-constraints towards diffusion processes in semiconductor manufacturing. Kopp et al. (2020) propose a novel rule-based dispatching concept validated in simulations that considers time-constraint criticality, setup times and general lot priorities. All in all, increasing high fidelity simulations improve the applicability of simulation based production control to complex job shops, the integration with near real-time data in digital twins however is still missing.

Heuristic dispatching control under time-constraints

Beyond the simulation validated approaches, further heuristic, rule based dispatching is proposed to increase time-constraint adherence. Chang & Chang (2012) regard time-constraints between wet etch and furnace areas that considers batching and a likelihood ratio for cycle time likelihood reduction for dispatching. A comparison of dispatching heuristics which affect time-constraint adherence is performed by Maleck & Eckert (2017) who conclude with a multi stage dispatching control policy related to Lee & Chen, et al. (2005) that considers tool availability, WIP inventory levels and time-constraints. Focusing on the implant work area Yang & Ke, et al. (2015) propose a rule based dispatching based on queuing theory that aims to increase same recipe rates which helps reducing setup times. The underlying motive is that in implant expiring time-constrained lots can skip the queue to be saved but come at the expense of large tuning beam and setup costs. The combination of heuristic dispatching with a neural network to select rule parameters by Li et al. (2012) and with binary integer programming in a two-machine flowshop is also explored by Ham & Lee, et al. (2011).

Alternatively, Wu & Lin, et al. (2010) use a markov decision process to derive control policies. They aim to minimize inventory holding and scrap costs, the latter stemming from time-constraint violations. Based on this well-known relationship they apply value iteration to obtain a simple control policy for a two-stage single product system. This approach is extended towards parallel processing systems (Wu & Lin, et al. 2012) as well as an upstream batching process (Wu & Cheng, et al. 2012). Then they improve their approach for multiple, different products (Wu & Chien, et al. 2016). In general these markov decision process derived dispatching rules decreased the scrap rate by up to 40% (Wu & Lin, et al. 2010) and up to 59% (Wu & Chien, et al. 2016). Due to the lack of applicability to larger systems owing to the modeling assumptions Wang & Ju (2021) use domain knowledge to decompose a complex

job shop into subsystems. Each of these is solved independently with policy or value iteration. Through iteration the overall policy is obtained.

Controlling Dispatching through time-constraint adherence prediction

Instead of controlling the dispatching to achieve multi-criteria optimization, several authors propose a novel production control approach for time-constrained complex job shops. The underlying rationale is to predict the probability of time-constraint adherence for all considered lots and control the gate keeping decision by excluding critical lots. Sadeghi et al. (2015) propose a probabilistic approach by representing the problem as a disjunctive graph and multiple times run a randomized list scheduling approach. Based on this randomization time-constraint adherence is estimated through the proportion of schedules that adhere to the time-constraint. Lima et al. (2017b), Lima et al. (2019) and Lima et al. (2021) extend this approach with intelligent sampling to improve probability estimation and reduce computational effort. One version of their algorithm is implemented into a decision support system to manage larger semiconductor fabs (Lima et al. 2017a) which identifies tool interruptions through shared recipes based machine grouping and time-constraint aggregation. These approaches aimed at applicability under real industrial conditions. Similarly, the author of this thesis in May & Maucher, et al. (2021) regards an entire semiconductor wafer fab as a complex job shop aiming to reduce time-constraint violation by predicting the adherence probability. Instead of sampling or randomization, the approach is based on a combination of historical transition time data and near real-time data to predict the transition time and time-constraint violation probability based on prediction intervals.

Dispatching summary

All in all, dispatching provides the largest lever to ensure time-constraint adherence in a complex job shop. Due to the stochastic nature and large effects of machine breakdowns and further stochastic processes, simulations and real-world data are frequently used to improve or validate the proposed algorithms. There is a plethora of dispatching rules that aims to reduce time-constraint violations but is based on the well known relationship between cycle time and inventory (Maleck & Eckert 2017). The results in general support the approaches, however, the regarded system size is often to small to control a complex job shop or a semiconductor wafer fab as its instantiation (Ham & Lee, et al. 2011). A promising approach with good results and applicability to larger, real industrial systems is the prediction of time-constraint adherence through a probabilistic (Lima et al. 2021) or auto correlated approach (May & Maucher, et al. 2021). With the advent of artificial intelligence and machine learning, novel approaches can used to improve the prediction accuracy.

3.1.4 Implementing learning based production control in job shops for time-constraints

On a planning level, Chien & Kuo, et al. (2020) use a neural network to improve their previous studies. They predict arrival rates and WIP inventory on a workgroup level to identify the influence of varying allocations. They aim to smooth WIP inventory and cycle time to increase the time-constraint adherence and overall throughput. On a dispatching level Schelthoff et al. (2022) investigate features to accurately predict waiting times in a real semiconductor fab. However, they do not study the influence on time-constraints or derive control actions to improve operations or time-constraint adherence. In contrast to Chakravorty & Nagarur (2020) who use an artificial neural network to control the gate keeping decision and restrict lots if the predicted transition time is larger than the time-limit. May & Behnen, et al. (2021) use a neural network and further multi-variate point estimators to predict the transition time. They extend this point estimator with a prediction interval to calculate the risk of violating individual time-constraints if lots are released at the gate keeping decision. With a similar quest May & Albers, et al. (2021) successfully predict future queues in front of equipment to derive transition time estimates and improve machine utilization.

A fundamentally different approach is presented by Wang & Hu, et al. (2020) who train a reinforcement learner to dispatch lots under time-constraints. However, their results are worse than the traditional approaches and suffer from high computational effort. Altenmüller et al. (2020) successfully train a deep reinforcement learning agent in an abstracted small complex job shop simulation with strict time-constraints. In a similarly sized simulation model Valet et al. (2022) control lot dispatching and maintenance to improve throughput and reduce scrap through time-constraint violation. The reinforcement learning approaches, however, are restricted to smaller size problems due to the slow speed of large scale simulations and high computational effort to train them.

All in all, learning based production control recently gained momentum with increased data availability and cheaper training. Simulation based reinforcement learning as of now lacks large scale applicability but provides promising results. Time-constraint violation prediction with the help of machine learning is showing good results, however, the interrelations of waiting times and queues as well as the transition from point estimators to uncertainty informed machine learning are challenging.

3.2 Research deficit

Considering time-constraints in all levels of the PPC taxonomy is a widely regarded area of research. Due to the dynamic nature of time-constraints and their fulfillment in a large real complex job shop in semiconductor manufacturing, production control on dispatching

level is most promising. To address this challenge, heuristics and mathematical models are most widely applied to deal with time-constraints in research and real fabs. The presented approaches are, however, not applicable in real-world scenarios as they typically heavily simplify the system, restrict the number of machines to only a handful and do not take into account transition heterogeneity in a real complex job shop. Therefore, results usually cannot be validated in a real-world semiconductor manufacturing environment. Thus, the real-world state-of-the-art method of choice is to employ human operators that individually support time-constraint adherence on local levels. Their decisions are based on logic and learning from the past. Hence, machine learning approaches are promising techniques to incorporate this decision making in an intelligent and automated way. Most pressing is, thus, the ability of production control to correctly handle gate keeping decisions with time-constraints to reduce scrap, increase yield and throughput on an operational level. While simulations have been used to train or derive such policies in small scale systems, their application or extension with digital twins is a generally promising field that has yet to be applied to intelligent production control of time-constrained complex job shops. The summary of the presented research work dealing with time-constraints in production planning and control for complex job shops is highlighted in the following, Table 3.1 reviews capacity planning and intelligent production control, Table 3.2 presents scheduling approaches and Table 3.3 regards dispatching approaches. The presented studies are clustered according to the dimensions introduced in Section 3.1. These are filled by the identified concept through the grounded theory based literature review. First, the modeling dimension can be associated to experiments, regarding data or a subsystem of a real wafer fab, to queuing systems, as a mathematical modeling, to MI(L)P, with a mixed integer linear or non-linear program, to disjunctive graphs, based on a graph theoretic understanding of the regarded system, or to markov decision processes. Second, the dimension solution approach makes use of traditional queuing theory or exact, for instance branch-and-bound methods. Alternatively, heuristics or meta-heuristics have been used which are nowadays complemented by machine learning approaches. Lastly, the objective dimension distinguishes the degree of interest in minimizing time constraint violations and similarly analyzes the application of time-constraint violation probability estimations. In total, the research deficit can therefore be summarized as follows and gives rise to the overarching research questions (RQ) introduced in Chapter 1.

1. Digital twins provide a valuable tool to improve production control but so far have not been applied to real world complex job shops and have not been used to control time-constraints in semiconductor manufacturing in real-time.

RQ1 How to use static and dynamic knowledge graph based production system replicas to support production planning and control with time-constraints?

2. Most existing approaches cannot effectively integrate real time real-world data and fail to properly ensure time-constraint adherence in complex job shops due to the dynamic nature and large problem size.

RQ2 How to use real-world real time data to avoid time-constraint violations with a data-based approach for production control for complex job shops?

3. Integrating intelligent, machine learning based approaches and their inherent uncertainty to quantify probabilities is a novel approach that has not yet been methodologically described properly and then applied to time-constraint complex job shops with different structures.

RQ3 How to enrich and extend machine learning algorithms to accurately capture the aleatoric and epistemic uncertainty in large-scale complex job shops when predicting time-constraint adherence?

4. Heuristics and simple, non intelligent production control policies are the norm and require frequent adaptations and human interventions to successfully control a complex job shop. These approaches are in particular not capable of acquiring knowledge and integrating it into decision making which results in the inability to effectively reduce time-constraint violations with production control.

RQ4 How to use long-term and real-time knowledge acquired within a factory to holistically reduce time-constraint violations with intelligent production control?

5. A real world applicability in semiconductor wafer fabs, with the specific aim to ensure time-constraint adherence, has rarely been considered for production control. A notable improvement over the industrial state-of-the-art of manual control for time-constraints has hardly been realized.

RQ5 How does the learning-based intelligent production control for complex job shop perform in ensuring time-constraint adherence in a real-world setting?

Table 3.1: Overview of relevant research on capacity planning and intelligent production control approaches for time-constrained complex job shops.

	Modeling					Solution					Object.	
Approach by	**Experiments**	**Queuing System**	**MI(L)P**	**Disjunctive graph**	**Markov decision process**	**Queuing Theory**	**Exact, branch-and-bound**	**Heuristic**	**Meta-Heuristic**	**Machine Learning**	**Min. time-constraint violation**	**Estimate violation probability**
Capacity planning (Section 3.1.2.1)												
Robinson & Giglio 1999	○	●	○	○	○	●	○	○	○	○	●	◐
Kitamura et al. 2006	○	●	○	○	○	●	○	○	○	○	●	◐
Ono et al. 2006	○	●	○	○	○	●	○	○	○	○	●	◐
Tu & Liou 2006	○	●	○	○	○	●	○	○	○	○	●	◐
Tu & Chen 2009a	○	●	○	○	○	●	○	○	○	○	●	◐
Tu & Chen 2009b	○	●	○	○	○	●	○	○	○	○	●	◐
Tu & Chen 2010	○	●	○	○	○	●	○	○	○	○	●	◐
Tu & Chen 2011	○	●	○	○	○	●	○	○	○	○	●	◐
Kuo et al. 2011	●	○	○	○	○	○	○	○	○	●	◐	○
Huang et al. 2011	●	○	○	○	○	○	○	●	○	○	◐	○
Pappert et al. 2016	●	○	○	○	○	○	○	●	○	○	◐	○
Chien & Kuo, et al. 2020	●	○	○	○	○	○	○	○	○	●	◐	○
Intelligent production control approaches (Section 3.1.4)												
Altenmüller et al. 2020	○	○	○	○	●	○	○	○	◐	●	●	○
Wang & Hu, et al. 2020	○	○	○	○	●	○	○	○	◐	●	◐	○
Chakravorty & Nagarur 2020	●	○	○	○	○	○	○	○	○	●	●	◐
May & Maucher, et al. 2021	●	○	○	◐	○	○	◐	○	○	◐	●	●
May & Behnen, et al. 2021	●	○	○	◐	○	○	◐	○	○	●	●	●
May & Albers, et al. 2021	●	○	○	○	○	○	○	○	○	●	◐	○
Schelthoff et al. 2022	●	○	○	○	○	○	○	○	○	●	○	○
Valet et al. 2022	◐	○	○	○	●	○	○	○	○	●	◐	○

Legend: ● considered ◐ partially considered ○ not considered

Table 3.2: Overview of relevant research on scheduling for time-constrained complex job shops.

	Modeling					Solution					Object.	
Approach by	**Experiments**	**Queuing System**	**MI(L)P**	**Disjunctive graph**	**Markov decision process**	**Queuing Theory**	**Exact, branch-and-bound**	**Heuristic**	**Meta-Heuristic**	**Machine Learning**	**Min. time-constraint violation**	**Estimate violation probability**
Scheduling (Section 3.1.2.2)												
Su 2003	○	○	●	○	○	○	○	●	○	○	◐	○
Sun et al. 2005	○	○	●	○	○	○	●	○	○	○	●	○
Bixby et al. 2006	○	○	●	○	○	○	●	○	○	○	◐	○
Chen & Yang 2006	○	○	●	○	○	○	●	○	○	○	◐	○
Ham & Raiford, et al. 2006	○	○	●	○	○	○	○	●	○	○	◐	○
Chien & Chen 2007	○	○	●	○	○	○	○	○	●	○	◐	○
Mason et al. 2007	○	○	●	○	○	○	○	○	●	○	◐	○
Klemmt & Horn, et al. 2008	○	○	●	○	○	○	●	○	●	○	◐	○
Yurtsever et al. 2009	○	○	●	○	○	○	○	●	○	○	◐	○
Kao et al. 2011	○	○	●	○	○	○	●	○	○	○	◐	○
Klemmt & Mönch 2012	○	○	●	○	○	○	●	○	○	○	●	○
Yugma et al. 2012	○	○	○	●	○	○	○	○	●	○	◐	○
Jia et al. 2013	○	○	●	○	○	○	○	○	●	○	◐	○
Jung et al. 2013	○	○	●	○	○	○	●	○	○	○	◐	○
Yu & Kim & Jung, et al. 2013	○	○	●	○	○	○	●	○	○	○	◐	○
Cho & Park, et al. 2014	○	○	●	○	○	○	●	○	○	○	◐	○
Jung et al. 2014	○	○	●	○	○	○	●	○	○	○	◐	○
Wang & Chien, et al. 2014	○	○	●	○	○	○	○	○	●	○	◐	○
Wang & Chien, et al. 2015	○	○	●	○	○	○	○	○	●	○	◐	○
An et al. 2016	○	○	●	○	○	○	●	○	○	○	◐	○
Wu & Zhao, et al. 2016	○	○	●	○	○	○	○	●	○	○	●	○
Kim & Lee 2017	○	○	●	○	○	○	●	○	○	○	◐	○
Maleck & Eckert 2017	○	○	●	○	○	○	●	○	○	○	◐	○
Maleck & Weigert, et al. 2017	○	○	●	○	○	○	●	○	○	○	◐	○
Yu & Kim & Lee 2017	○	○	●	○	○	○	○	●	○	○	◐	○
Zhou & Wu 2017	○	○	●	○	○	○	○	○	●	○	◐	○
Maleck & Nieke & Bock & Pabst & Stehli 2018	○	○	●	○	○	○	●	○	○	○	◐	○
Wang & Srivathsan, et al. 2018	○	○	●	○	○	○	○	●	○	○	◐	○
Maleck & Nieke & Bock & Pabst & Schulze, et al. 2019	○	○	●	○	○	○	●	○	○	○	◐	○
Nattaf et al. 2019	○	○	●	○	○	○	○	○	●	○	●	○
Perraudat et al. 2019	●	○	○	○	○	○	○	●	○	○	◐	○
Zhou & Lin, et al. 2019	○	○	●	○	○	○	○	○	●	○	●	○
Lee 2020	○	○	●	○	○	○	○	○	●	○	◐	○

Legend: ● considered ◐ partially considered ○ not considered

Table 3.3: Overview of relevant research on dispatching for time-constrained complex job shops.

	Modeling					Solution					Object.	
Approach by	**Experiments**	**Queuing System**	**MI(L)P**	**Disjunctive graph**	**Markov decision process**	**Queuing Theory**	**Exact, branch-and-bound**	**Heuristic**	**Meta-Heuristic**	**Machine Learning**	**Min. time-constraint violation**	**Estimate violation probability**
Dispatching (Section 3.1.3)												
Scholl & Domaschke 2000	●	○	○	○	○	○	○	●	○	○	◐	○
Lee & Chen, et al. 2005	●	○	○	○	○	○	○	●	○	○	●	○
Tu & Chen & Liu 2010	●	○	○	○	○	○	○	●	○	○	◐	○
Wu & Lin, et al. 2010	○	○	○	○	●	○	○	○	●	○	◐	○
Ham & Lee, et al. 2011	○	○	●	○	○	○	○	●	○	○	◐	○
Chang & Chang 2012	○	○	●	○	○	○	○	●	○	○	◐	○
Li et al. 2012	○	○	●	○	○	○	○	●	○	◐	●	○
Wu & Cheng, et al. 2012	○	○	○	○	●	○	○	○	●	○	◐	○
Wu & Lin, et al. 2012	○	○	○	○	●	○	○	○	●	○	◐	○
Kobayashi et al. 2013	●	○	○	○	○	○	○	●	○	○	◐	○
Toyoshima et al. 2013	●	○	○	○	○	○	○	●	○	○	◐	○
Ciccullo et al. 2014	●	○	○	○	○	○	○	●	○	○	◐	○
Arima et al. 2015	●	○	○	○	○	○	○	●	○	○	◐	○
Sadeghi et al. 2015	○	○	○	●	○	○	○	●	○	○	●	●
Yang & Ke, et al. 2015	○	●	○	○	○	○	○	●	○	○	◐	○
Wu & Chien, et al. 2016	○	○	○	○	●	○	○	○	●	○	◐	○
Zhang et al. 2016	●	○	○	○	○	○	○	●	○	○	●	○
Lima et al. 2017a	○	○	○	●	○	○	○	●	○	○	●	●
Lima et al. 2017b	○	○	○	●	○	○	○	●	○	○	●	●
Lima et al. 2019	○	○	○	●	○	○	○	●	○	○	●	●
Kopp et al. 2020	●	○	○	○	○	○	○	●	○	○	●	○
Pirovano et al. 2020	●	○	○	○	○	○	○	●	○	○	◐	○
Lima et al. 2021	○	○	○	●	○	○	○	●	○	○	●	●
Wang & Ju 2021	○	○	○	○	●	○	○	○	●	○	◐	○

Legend: ● considered ◐ partially considered ○ not considered

4 Intelligent Production Control for time-contrained complex job shops

Within this chapter the methodological approach is developed in order to achieve the objectives set out in Section 1.3 and to address the research deficit as derived in Section 3.2. The overarching goal is to develop a novel production control method for complex job shops that can improve time-constraint adherence. The investigation is based on the implementation of a data-driven algorithm for making the gate keeping decision in front of time-constraints. In order to implement this approach, traditional time series predictors, machine learning algorithms and their uncertainty are regarded to improve the applicability of uncertainty driven intelligent decision making in manufacturing. Furthermore, the integration of a knowledge graph based digital twin is analyzed. The overall structure can be seen in Figure 4.1 which introduces the methodological overall approach. Possibly, the result of this research is an applicable method to derive a data-driven, intelligent complex job shop production control for time-constraint gate keeping decisions to avoid manual effort and increase the time-constraint adherence. Owing to the ever increasing complexity in manufacturing, the transferability of the approach and of some of the developed tools to different complex job shops and production control methods motivates this research.

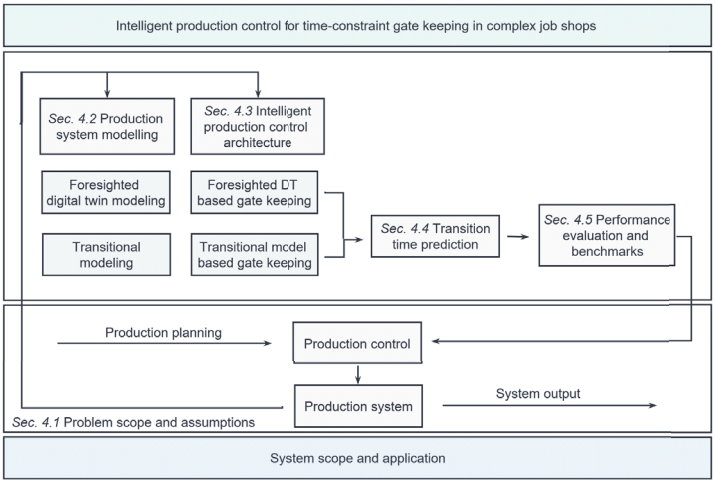

Figure 4.1: Methodological approach to obtain the intelligent production for time-constraint gate keeping in complex job shops.

The general problem and its scope and assumptions are framed in Section 4.1 to facilitate the overall problem solving approach. Then, the description of the proposed method is split into two parts. First, Section 4.2 introduces the modeling of the production system and an instantiation as a digital twin based on a discrete event simulation answering RQ1. Second, the general architecture to implement this intelligent production control for time-constraint gate keeping in complex job shops is described in Section 4.3 to address RQ2. Next, the approaches used to predict transition times and derive time-constraint adherence probabilities based on uncertainty quantification are introduced in Section 4.4 to present a solution to RQ3. The evaluation of the proposed transition time predictors and the comparison with benchmarks is presented in Section 4.5 and hence addresses RQ4. The application to a real-world complex job shop in a semiconductor wafer fab is later introduced in Chapter 5 to answer the final research question 5.

During previous research of the author of this thesis several aspects and ideas that this thesis is based on have been introduced. Most notably, the following research is used. Firstly, the conceptualization of a foresighted digital twin is introduced by May & Overbeck, et al. (2021). Secondly, one approach to the production system modeling with the ontology based OntologySim as a foresighted digital twin is published in the paper by May & Kiefer & Kuhnle & Lanza (2022). Thirdly, the understanding of the real-world complex job shop and inventory predictions are introduced by May & Albers, et al. (2021). Additionally, a uni-variate prediction approach is presented in the work of May & Maucher, et al. (2021) while a multi-variate approach is presented by May & Behnen, et al. (2021).

4.1 Problem scope and assumptions

The general scope of this work is narrowed down in Chapter 1 and explained in Chapter 2. Based on the identified research deficit as outlined in Chapter 3, the scope can be specified with the following assumptions:

1. A large complex job shop is regarded which is represented as an entire semiconductor wafer fab as the research object. Within this manufacturing system lots denote products that need to be processed in a reentrant manner. Transitions between processing equipment may or may not be time-constrained as identified through real-time data from operations. Production orders arrive according to the production program set outside the system.

2. In the PPC hierarchy dispatching is regarded insofar as that the gate keeping decision in front of equipment whether or not a particular time-constrained lot is processed is controlled. A single decision making algorithm for the entire manufacturing system is regarded.

3. Due to the differing criticality of time-constraints, the decision making algorithm can make use of a hierarchical algorithm that treats more critical time-constraints differently and with more computational effort than uncritical time-constraints.

4. Real-time, real-world production data represents the real-system so that decisions are based on the actual state of the complex job shop. Through regarding different time-snippets of the systems over multiple years a generalization can be achieved.

5. For validation purposes the proposed algorithms are implemented and fit during one part of the time-frame and then evaluated on an unknown, unseen part of the time-frame of the regarded production system. The validation takes place with the comparison of past behavior as observed or proposed by the developed algorithm. Nevertheless, the actual implementation and roll-out of this algorithm in the real semiconductor wafer fab is not part of this work.

6. To ensure possible application in the ever changing setup of a complex job shop, frequent tuning and training of the algorithm will be necessary. The proposed algorithm implements and includes these updates for short-term changes within the time-frames that span several months each. However, for longer-term application regular tuning and training of the algorithm will be necessary. Due to the regarded different time-frames, this behavior is already mimicked but not regarded explicitly as a part of this work.

4.2 Modeling the production system

A prerequisite to the following steps of the introduced model lies in understanding the coherent modeling approach. In fact, two modeling approaches are required: Firstly, to possibly evaluate production control decisions before enacting them, a foresighted digital twin is required. Secondly, to frame the decision making framework for the gate keeping decision of time-constraints a novel approach is required. For the former there are multiple reasons a discrete event simulation model should be used. Real-world dynamics have to be implemented and the behavior of the system under decision making should be studied. To create this foresighted digital twin a homogeneous data model in both real-world manufacturing system and the digital twin should be used which is flexible and extendable. A modeling approach that facilitates the decision making framework is required to implement the intelligent production control in the subsequent chapters.

The organization of this section follows the visual presentation in Figure 4.2. The most relevant system elements of this complex job shop are described in Section 4.2.1. Subsequently a simulation and digital twin model is derived which is presented in Section 4.2.2. Based on this system modeling and to enable the intelligent production control for time-constraint gate

keeping decisions, the transitional modeling approach is derived. Section 4.2.3 presents this approach. Neither the proposed simulation and digital twin, nor the transitional modeling approach aim at completeness for applicability on all levels but form the prerequisites for the following decision making as introduced in later chapters.

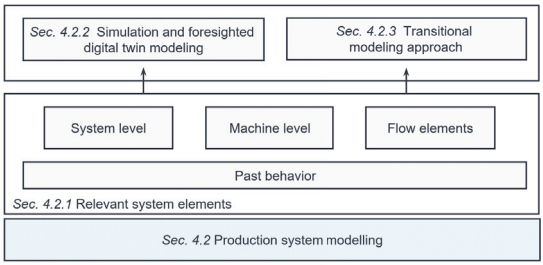

Figure 4.2: Organization of Section 4.2.

4.2.1 Relevant system elements

The aim of this research is to provide a methodological approach and solution to gate keeping production control for time-constraints in complex jobs shops. While applicability in real-world manufacturing systems is required the approach should be as generic as possible. During the manufacturing process this gate keeping decision can be triggered whenever a possible time-constraint lot may be selected to start processing. The relevant elements that need to be considered are based on industrial partner discussion and extended with system elements typically used in the identified literature review. Thus, in order to capture a complex job shop with time-constraints the following system elements need to be considered.

System level influences

First, the setting and external influences need to be considered on a systemic level. The overall system setting suffers from large and increasing competition, so that operational excellence is paramount. Therefore, within a complex job shop, possibly with thousands of machines as in the regarded semiconductor wafer fab, utilizing this capital intensive equipment should not be compromised and aggressive due dates need to be met. This leads to a large fluctuation in demand both on an individual level and on an aggregated demand figure. The result is an environment characterized by highly dynamic behavior and order releases into the manufacturing system, in other words systems are operated with dynamical operating points. Hence, product release is a key element that needs to be modeled within the approach.

A similar speed of change happens on the shopfloor as material flow equipment can be exchanged, upgraded or is under maintenance. In a complex job shop such temporary or longer-term changes do not disrupt the production but still need to be reflected in the modeling approach. This results in a need to coherently abstract the modeling approach from static layout and planning decisions that are usually the foundation of a manufacturing system model to a dynamic approach.

Machine level elements

On a machine and equipment level frequent and stochastic breakdowns have a decisive effect on the production system. Therefore, machines must be modeled as the central elements of the production system. In contrast to static modeling these stochastic breakdown dynamics have to be considered and modeled within the approach. Not only the breakdown dynamics are stochastic and have a large spread but also the processing time in a complex job shop can vary greatly. The processing time depends on product specific requirements, i.e. process complexity for this particular product, on the equipment condition and on the random behavior and quality rate during the processing. Thus, these stochastic process times need to be considered. Furthermore, setup times depend on sequences, i.e. the time required to setup the machine for a particular product depends on both the product and the previous product on this machine. Such sequence dependent setup times need to be modeled as important elements.

A larger effect can be attributed to machine dedications as certain machines are restricted to performing certain processes and due to the technological complexity in complex job shops. Therefore, machine capabilities must be modeled within the approach. Due to the different types of machines and their differing behavior, single job, batching, continuous processes and alternatives must be modeled within the machine. Additionally, a control system in front of the machines, the so called gate keeping decision, has to be modeled as a key element to sustain both the variability created by machine dedications and breakdowns as well as the gate keeping decision that is to be implemented during this thesis.

Material and process flow elements

A decisive driver for the complexity of complex job shops is the heterogeneity of products and their general modeling. Therefore, products and variants need to be flexibly modeled and have to contain required processes and interrelations such as time-constraints. The result is a product that comes with required processes that can be matched with the capabilities on the machine level. These products' structured processes define the route that the products are taking through the complex job shop. In semiconductor manufacturing this includes reentrant flow, so that several hundred process steps and recurrently visiting machines or machine

groups have to be modeled. Time-constraints need to be modeled and associated with the respective processes, so that coherent modeling is ensured.

Overall, the property of irregular, non linear material flow is an observable fact for complex job shops that needs to be modeled. The underlying rationale is that the knowledge about availabilities and operational capabilities of machines coupled with required processes from products alone are insufficient to describe this behavior. At the core of this orchestration is the production control from the PPC hierarchy. Therefore, the production control level and information shall be directly or implicitly included in the modeling of the system.

Past behavior

As the dynamics in a complex job shop unfold over time many indirect relations and influences can only be implemented and only become obvious if long term behavior and short term behavior, in other words the recent behavior of the system, is regarded and can be used to model the system. This element of a past behavior is a fundamental difference to standard modeling approaches. The way of integrating this modeling as well as the degree and length of past behavior considered, nevertheless, is regarded in the following research.

4.2.2 Simulation and foresighted digital twin modeling

At the core of the foresighted digital twin to be construct lies a simulation (May & Overbeck, et al. 2021) which is a discrete event simulation (DES) model for the scope of an entire production system. To obtain the flexibility and changeability in the simulation model, necessary to map to a complex job shop, it is based on an ontology and knowledge graph as introduced in Section 4.2.2.1. For a foresighted digital twin the interconnection to the real-world system's real-time data and behavior is necessary which uses a state representation in form of a knowledge graph as explained in Section 4.2.2.2.

4.2.2.1 Ontology-based production simulation

In order to implement the relevant system elements into the simulation and achieve a high validity, the approach is based on the OntologySim introduced by May & Kiefer & Kuhnle & Lanza (2022). The elements that are modeled within the simulation can be categorized into products, resources and events. Instead of the traditional wording of orders, products is used to clearly denote the reentrant flow and enable a decoupling of customer orders (of potentially multiple products) and manufacturing internal lots regarded in the real-world semiconductor manufacturing use-case. While resources follow the traditional system modeling and naming approach, events constitute a novel group elements. To create such a flexible simulation and following the OntologySim approach events are defined to coherently describe the system

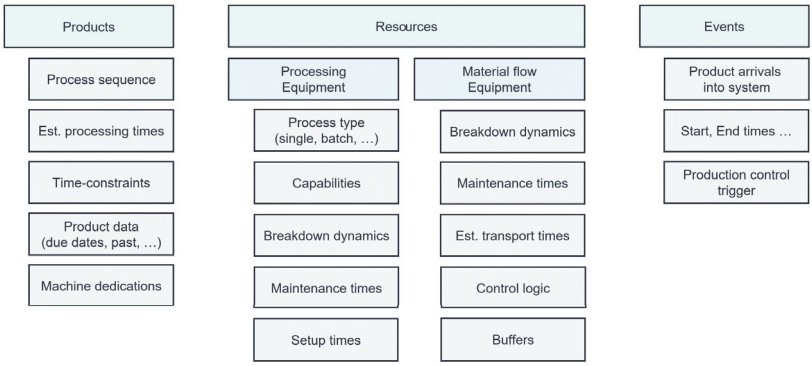

Figure 4.3: Categorization of system elements in the simulation.

and its state. Figure 4.3 illustrates these types of elements and the respective content of these elements which are introduced in the following.

Product definition

In this generalized approach a product is defined by its route or process sequence, the current state within this sequence and associated events, time-constraints within the route and general information. Through the state and general information about location, quality information and a unique identifier the product is instantiated from the abstract route definition. A route is defined by a sequence of processes that need to be undergone during the manufacturing. To account for complex production plans that allow for different paths through this sequence a graph based sequence structure with linked predecessors and successors is kept. Within each graph element a process is described including information about (estimated) processing times and time-constraints. The latter is stored as a beginning and ending process associated with a time limit. Doing so permits the realization of all types of time-constraints. As the product is part of the overall ontology, or of instantiated the overall knowledge graph, events are also associated with the product. While not regarded in this use-case, merging and unmerging of products, for instance in disassembly, can be achieved with this flexible knowledge graph based approach. This is visualized in Figure 4.4 with product and events as overarching concepts as well as datatypes highlighted. Clearly, the interrelations between processes and time-constraints despite being part of different system elements, as introduced before, can be seen. On the lowest level concrete data types, such as arrays for process information, can be seen.

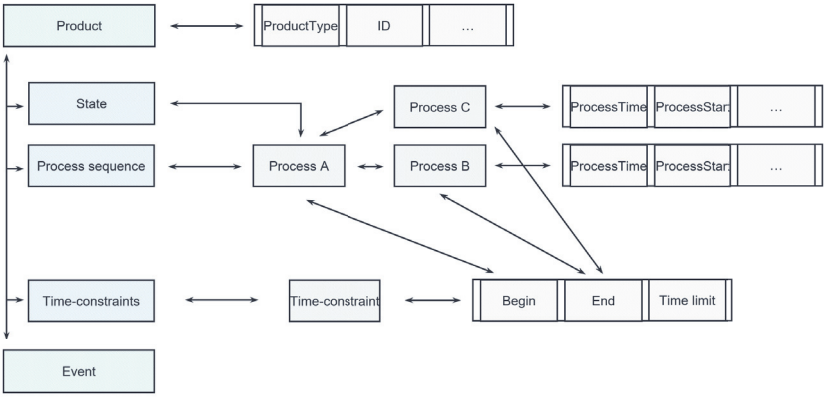

Figure 4.4: Concept of a simplified product modeled in the simulation based on May & Kiefer & Kuhnle & Lanza (2022).

Resource definition

As processes change the properties of products during production they resemble the central element in a production system. Additionally, time-constraints restrict the transition time between the end and beginning of two processes. The concept of such a processing resource or machine is described first and visualized in Figure 4.5. The underlying motive is the representation of the material flow within such a processing resource as queues, based on May & Kiefer & Kuhnle & Lanza (2022). Products can be in a buffer state, both in- and outbound which may even be the same queue, or in the production process (prod_queue). Each processing resource can perform the process according to its capabilities and technical refinements for these processes such as maximum temperatures or available production chamber space. Here, through implicit couplings of the positions in a production queue, batch processing can be realized. Likewise, the machine itself can be in a down state with different down types such as planned, unplanned, during repair or waiting for maintenance, in an up and running or idling and setup state. This behavior and the end of a production process is triggered, as in every discrete event simulation, through the event list. However, these events are modeled and connected to the resource concept and product concept, so that the interrelation between resource, product and event, for instance in the end of processing where a product is transferred from a buffer position the the production position, is accurately reflected.

Note that each of theses subentities are modeled based on the regarded production system, so that the detailed information about the types of products, downtimes, setups etc. is similarly

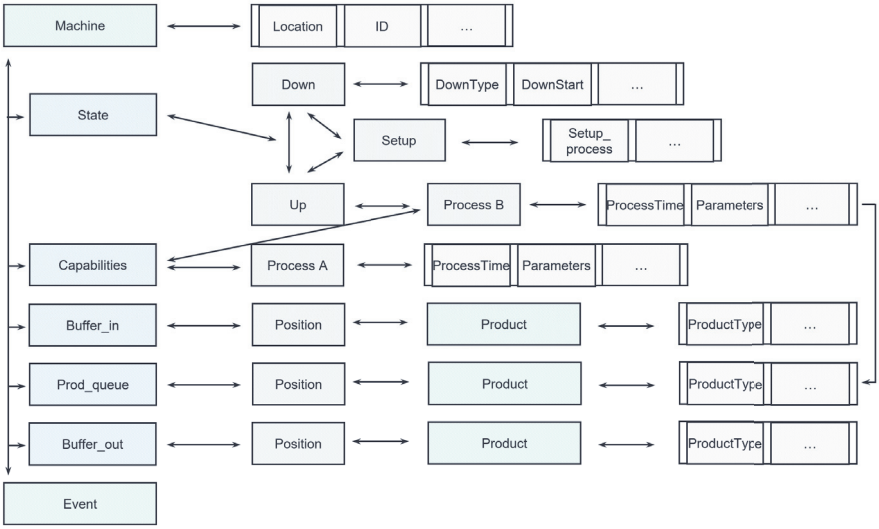

Figure 4.5: Concept of a simplified processing resource modeled in the simulation based on May & Kiefer & Kuhnle & Lanza (2022).

saved in the ontology described knowledge graph to obtain a single source of truth during the simulation. Therefore, both deterministic known and stochastic behavior can be tied to this. Each resource has additional data such as location or neighborhood information and a unique identifier. Through the ontology and knowledge graph flexibility this processing resource concept can be easily extended to individual use-case need. The required control logic for product selection to the next process can incorporate a simple first-in-first-out heuristic or be extracted through the event handling as described in the following.

Secondly, material flow equipment is described. The general setup of a resource consisting of a queue is kept. For conveyor style material flow equipment this behavior is an accurate reflection, while for automated guided vehicles and individual transport, the queue length is restricted to the capacity and the location becomes flexible. Note that a relationship between entities from zero to $n \in \mathbb{N}$ are possible due to the ontology based knowledge graph. Similarly to processing equipment breakdowns and maintenance are modeled. Average speeds or real observed speeds can be tied to the respective equipment. Through location and route information the estimated transportation time can be obtained. However, pure transportation time makes only a small percentage of the overall transition time between two processes as waiting, sorting or rerouting have a much larger effect. To mimic real-world complex job shop

behavior and avoid deadlocks material flow equipment can pre-reserve queue positions if enacted by the production control logic. The control logic is integrated through an interface and can query all the information available within the knowledge graph.

Event definition

To recreate the dynamics of a real manufacturing system within the simulation, events are used. An event is stored within the ontology and can be a past, present or future event. To attain real-world comparability future events such as future breakdowns or processing times are restricted and not observable for production control or any simulation controller. Every event is characterized by its starting time, its (expected) duration and type as well as a unique identifier and the interconnection to the relevant products and resources. Based on the approach from May & Kiefer & Kuhnle & Lanza (2022) the actual state of each resource can be referred to the events and if past events are stored past states of products and resources can be recreated. Indirectly every event is interconnected with products and resources but it is not required that both resource and product are connected, e.g. for machine breakdowns or time-constraint violations. Figure 4.6 visualizes the structure of an event.

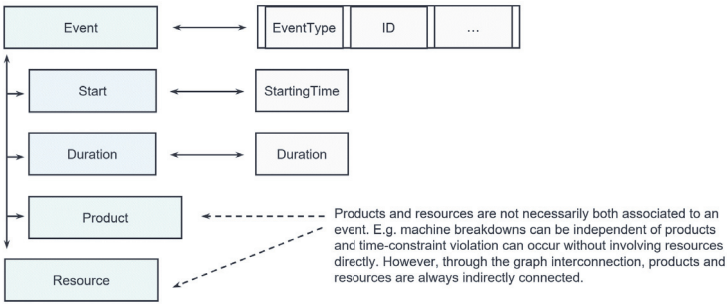

Figure 4.6: Concept of events modeled in the simulation based on May & Kiefer & Kuhnle & Lanza (2022).

Events are created either from scratch during the simulation initialization and the production control that assigns products to material flow and processing equipment or from past events. Exemplary the processing of a job at a machine is explained. First at the end of the maintenance period the next machine breakdown is sampled from the mean time to failure statistics and a future event is created. At the same time a particular product from the buffer requiring a certain process is selected and the following two events are created: Firstly, the setup event starting at the same time and requiring the setup duration. Secondly, the beginning of processing after the setup for the process duration. In case the before sampled breakdown

event occurs before the end of setup and processing events these events are interrupted for the breakdown duration. After the setup event has elapsed it is transferred to the past events.

Compact ontologies and small knowledge graphs have significant performance advantages (Lamy 2016). Thus, past events are only made available one directionally. Any other events and entities in the simulation ontology and instantiated knowledge graph are bidirectionally connected to enable inferring knowledge, implementing flexible changes and provide direct access. This tremendously reduces computational effort as only fewer, new events have to be iterated and queried (May & Kiefer & Kuhnle & Lanza 2022). Besides computational efficiency this flexibility motivates the usage of the ontology based approach as opposed to traditional relational or object oriented modeling.

Reasoning and initialization

A valid configuration is necessary to initialize the simulation. Thus, the first step starts with a reasoning over the provided system description, which may be a knowledge graph in a standard exchange format such as owl. The procedure is based on the implementation of the OntologySim which was introduced by May & Kiefer & Kuhnle & Lanza (2022). Then, a wrapper class approach is implemented in which all elements in the knowledge graph are made directly accessible through the programming of a wrapper class in the simulation. This tremendously reduces computational effort during the simulation, as fewer queries to access the entities are necessary. Nevertheless, the possibility to query and do reasoning on the knowledge graph itself remains.

For the initialization initial events have to be created. For instance breakdowns are sampled and current resource states determined. Through the knowledge graph based approach it is possible to warm start the situation and alter it during runtime. In other words products can already be placed in the manufacturing environment, for instance in buffers and processes and real-world breakdowns can be dynamically implemented. If warmstarts are used process starting times may be prescribed or derived from the system state.

Simulation process

During the simulation run all events are connected to an event lists, as visually introduced in Figure 4.7. The current simulation time is in the knowledge graph and referred to from the event list. Past events are connected unilaterally but not queried. During one time-step all currently open events are queried if their starting date plus duration is equal or smaller than the current time. These events are then finished, transferred to past events and possibly novel events created. In this example, after event 1 has just passed, it was transferred to

past events. Then, two novel events are created, the first one being event 3 for a setup and secondly the future event 5 for processing. During every discrete event time step production control can be triggered. Then the time is incrementally increased to the start of the next event or the end of any current event. Start times or durations of events can be increased or changed by other events, e.g. an unscheduled breakdown needs maintenance and the end of a process is deferred by that required time.

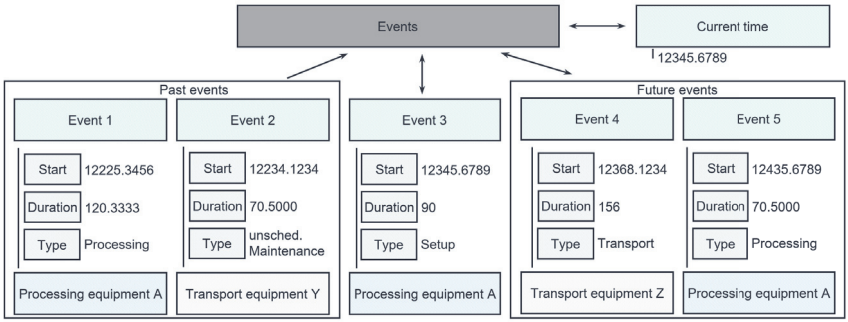

Figure 4.7: Events during an exemplary simulation run based on May & Kiefer & Kuhnle & Lanza (2022).

Production control can be triggered based on individual events, through an inserted production control event corresponding to a timely trigger, as a general control logic that is queried at any event or through manual interventions. Manual interventions should be avoided if simulation studies are performed as there is a great risk of manually inserting errors.

Product release into the system and order dispatching i.e. transport to machines and gate keeping decisions in front of the process equipment constitute possible production control. Order release is contacted at any event and can insert a novel event of releasing an order into the system. Either real production plans or statistical estimations can be used. Order dispatching is based on the prescribed process sequence for each product. It comprises transport decisions between different process equipment but cannot freely assign products to machines. Rule-based behavior from the regarded use case is implemented although extensions for the inclusions of reinforcement learning dispatching, as outlined by Kuhnle et al. (2022), are prepared. At the end of each processing step, in other words once the processing event on a processing equipment is finished, the next product to be processed needs to be selected. The selection can be done rule-based as in real-world complex job shops. During this the gate keeping decision for each lot in the buffer has to be taken which is interfaced to the production control to be developed.

4.2.2.2 Foresighted digital twin based on knowledge graphs

To glimpse into one or several likely paths the dynamics of a production system take from a certain point in time the application of a foresighted digital twin is used (May & Overbeck, et al. 2021). This foresighted digital twin follows the scheme outlined in Section 2.4.2. For the knowledge graph based instantiation real-world real-time system data is aggregated into a graph and includes past events and current events starting time and type. To instantiate a foresighted digital twin it has to be enhanced with production control transfer, the possibility to track KPIs in the used simulation and data imputation to fill in missing information from the real system. Reasoning over the knowledge graph can support this data imputation. Finally, a deep copy of this real system representing knowledge graph is made and used within the simulation model to generate foresight as illustrated in Figure 4.8.

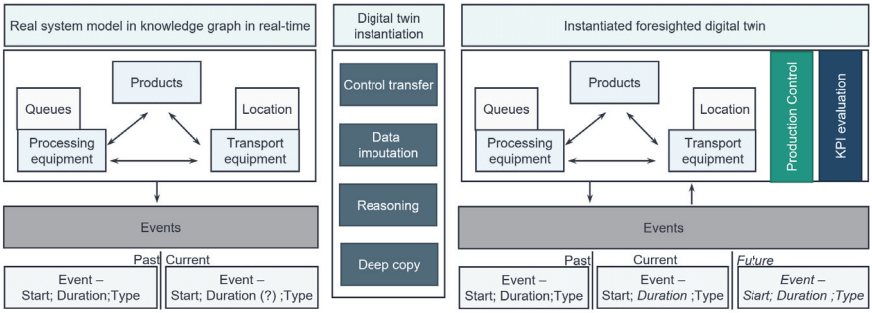

Figure 4.8: Product process flow chart with the regarded PPC decisions and the focus on time-constraint gate keeping decisions.

Deep copy

The so created instantiated ontology of the current real-system status, as shown in Figure 4.8, cannot be altered as it is interlinked with the real system. Instead a deep copy of this knowledge graph is to be made. This deep copy is the initial state of the digital twin and represents the current status. Before being able to use this current status, the deep copy needs to be completed to be capable of running the digital twin and obtaining foresight.

Control transfer

If digital twins are not used to train and improve production control reflecting the actual production control with high accuracy is important (Overbeck & Rose, et al. 2022). On the production system level Programmable Logic Controller (PLC) control is abstracted into the machine behavior as simulated. Releasing lots as the order release into the regarded system

is based on a prescribed process resulting in an order release schedule. Therefore, the short-term schedule can be transferred easily into the digital twin as the foresight period hardly exceeds this time horizon. Thus, the main complexity lies in transferring the scheduling and dispatching control logic into the digital twin.

Scheduling is performed on a time horizon of hours to shifts, so that at the time of instantiation a partial schedule is still present that can be copied into the digital twin. As the computational effort for these optimizations is high, and results are only obtained after waiting time, the inclusion of a scheduling mechanism into the foresighted digital twin would massively hinder the foresight as valuable time is wasted. In a complex job shop a great part of the initially proposed schedule is anyhow overthrown by real-time dispatching as breakdowns, time-constraint violations and the actual behavior does not permit the direct implementation of the schedules. Therefore, for a foresighted digital twin relying on the dispatching for real-time decision making is sufficient.

Short-term, real-time decision are taken by dispatching and sequencing in front of machine queues which includes the gate keeping decision. Dispatching is currently controlled by priority rules in real-world complex job shops (Kuhnle et al. 2022). Therefore, the knowledge about the priority rule used can obtained from interactions with workers. Due to their simplicity, they can be easily transferred and implement. The main challenge lies in the transfer of sequencing algorithms in front of processing equipment as heterogeneous priority rules, machine dedications, and their temporal inconsistency impede automation (Waschneck et al. 2016). Therefore, if the control transfer is not easily manageable, the following simplification can be used.

Instead of regarding groups of similar equipment and shared queues with machine dedications each equipment can be considered separately in a routing allocation problem. For a product the route is expected to include priorities for equipment or random sampling is assumed. In case of breakdowns the machine dedication information but not necessarily the exact algorithm for redistribution needs to be known as human interaction cannot be fully included in the digital twin ex ante.

Data imputation and reasoning

To deal with incomplete data in the real-systems database missing data has to imputed. Consider the value x_i being unknown but related to x_j through temporal, averaged or material (i.e. physical) relationship. When regarding KPIs intrinsic material relations, for instance through a defining graph are well known (May & Fang, et al. 2022), these should be used to impute missing data. The imputation algorithm is then described through pseudocode in Algorithm 1. It can be used to obtain the imputed data x_i^{impute} that shall not exceed an

expected small derivation of ϵ from the original value based on the L^2 norm that implements a quadratic distance function in a two dimensional euclidean space.

Algorithm 1 Imputation Algorithm

Require: x_i is unknown, x_j for $j \neq i, j \in J$ known with $i \cup J = \{0, 1, 2, 3, ..., n\}$ with $n > 5 \in \mathbf{N}$

Ensure: Minimize ϵ with $(x_i^{impute} - x_i)^2 \leq \epsilon$

$x_k^{impute} \leftarrow \frac{x_{k-1}+x_{k+1}}{2} \quad \forall k-1, k, k+1 \neq i$

$\epsilon_{temporal} = \frac{1}{n-5} \sum_{k=\mathbb{N}k-1,k,k+1\neq i}^{n-1} (x_{k-1}^{impute} + x_k^{impute} + x_{k+1}^{impute} - x_{k-1} - x_k - x_{k+1})^2$

$\epsilon_{averaged} = \frac{1}{n-1} \sum_{k \in J} \left[(\frac{1}{n-1} \sum_{h \in J} (x_h) - x_k)^2 \right]$

let $\epsilon_{material}$ denote the error from an approx. known relationship of x_i **and** the known y_i

if $\epsilon_{material} \leq \epsilon_{averaged}$ **and** $\epsilon_{material} \leq \epsilon_{temporal}$ **then**

$x_i \leftarrow$ calculated according to relationship with y_i

else if $\epsilon_{temporal} \leq \epsilon_{averaged}$ **and** $\epsilon_{temporal} \leq \epsilon_{material}$ **then**

$x_i \leftarrow \frac{x_{i-1}+x_{i+1}}{2}$

else

$x_i \leftarrow \frac{1}{n-1} \sum_{k \in J} x_k$

end if

In order to deal with heterogeneously missing data and different data point, this imputation algorithm favors simple and computationally efficient approaches. Thus, for temporal relationships a linear relationship between x_{i-1}, x_i, x_{i+1} is regarded and the mean average quadratic deviation used as an error. As the main information, necessary to obtain, is limited and restricted to times, failures etc. this restriction, for the sake of computational speed, is acceptable. If values are sampled from a random distribution without auto-correlation or remain constant, the average quadratic deviation from the mean is used. For many processes in a production system inherent relationships are known, so that these can be used to impute the missing values. Based on the temporal locally lowest estimated error the missing value is imputed.

Additionally, reasoning over the knowledge graph is necessary to complete information not directly contained in or imputed from the real system information. The requirement for reasoning is the availability of an ontology or a filled knowledge graph. Most importantly, future events have to be associated to the event lists. Let $e_{i,m} \in E$ denote event e_i associated with equipment $m \in M$. Then $a_{e_{i,j}}$ denotes the start of this event, $b_{e_{i,j}}$ denotes the type $b \in B$ and $d_{e_{i,j}}$ the duration. Each equipment m has the following three event lists: $E_{p,m}$ for past events, $E_{c,m}$ for current events and $E_{f,m}$ for future events. Reasoning creates the initial future events and potentially missing duration. Algorithm 2 introduces the reasoning algorithm to obtain missing values and events for breakdowns. The underlying approach is to iterate over existing events from the past and current list to generate the missing future events, for instance to

create a maintenance event after the expected time to failure. Likewise, current breakdowns are reasoned to create an event for normal operation resumption. All other values that are acquired from the knowledge graph follow the same scheme.

Algorithm 2 Event reasoning Algorithm

Require: $a_{e_{i,j}}, b_{e_{i,j}}$ are known for all $e_{i,j} \in \bigcup_{j \in M} \left[E_{p,j} \cup E_{c,j}\right]$ and $d_{e_{i,j}}$ is known for all $e_{i,j} \in \bigcup_{j \in M} E_{p,j}$

Ensure: missing $d_{e_{i,j}}$ and $e_{i,j}$ are known

```
for m ∈ M do
  if ∃b_{e_{i,m}} ∈ B_{breakdowns} ∀e_{i,m} then
    if b_{e_{i,m}} ∉ B_{breakdowns} ∀e_{i,m} ∈ E_{c,m} then
      j = max i   ∀b_{e_{i,m}} ∈ B_{breakdowns}
      if ∃e_{i,m} ∀b_{e_{i,m}} ∈ B_{breakdowns} then
        create e_{j*,m} with b_{e_{j*,m}} ∈ B_{breakdowns}, a_{e_{j*,m}} = a_{e_{j,m}} + d_{e_{j,m}} +
        MTBF_{sampled,m}, d_{e_{j*,m}} = MTTR_{sampled,m}
      end if
    else
      d_{e_{j,m}} = MTTR_{sampled,m}
      if d_{e_{j,m}} ≤ now then
        d_{e_{j,m}} = now
      end if
      create e_{j*,m} with b_{e_{j*,m}} ∈ B_{breakdowns}, a_{e_{j*,m}} = a_{e_{j,m}} + d_{e_{j,m}} + MTBF_{sampled,m}, d_{e_{j*,m}} =
      MTTR_{sampled,m}
    end if
  else
    create e_{i*,m} with b_{e_{i*,m}} ∈ B_{breakdowns}, a_{e_{i*,m}} = 0 + MTBF_{sampled,m}, d_{e_{j*,m}} =
    MTTR_{sampled,m}
  end if
end for
```

Running the foresighted digital twin

This digital twin contains the current status amended with missing and reasonable values and events. Given the discrete event simulation capability this simulation coupled with the transferred production control policies can start a simulation to create foresight. By changing production control policies or taking different production control policy decisions in copies of this digital twin the short term behavior can be analyzed as the KPIs are holistically stored. One possible implementation of this approach is presented and discussed in May & Kiefer & Kuhnle & Lanza (2022).

4.2.3 Transitional modeling approach

In contrast to the foresighted digital twin based modeling approach the transitional approach follows a different rationale. Not an executable system representation of the entire system is used but a subaggregation model as introduced in the following is implemented. The transitional modeling approach based on the work of May & Maucher, et al. (2021) regards the disjunctive set of processing and transporting operations. Processing creates value and takes place in processing equipment which is regarded as stationary. Any time spent between processes is regarded as non value creating and may be spend on a transport equipment or in a buffer. Due to the material flow complexity in a complex job shop the exact route, involved transport equipment, buffers and time required may vary heavily and are a-priori unknown. Therefore, the short term static processing equipment is modeled as vertices V in a graph $G = (V, E)$. The possible transition of any product through transporting and waiting is modeled as an edge $e \in E$ in the graph G. While Figure 4.9 presents this transition and graph structure analogously to the actual layout, G is a general graph and does not rely on this spatial representation.

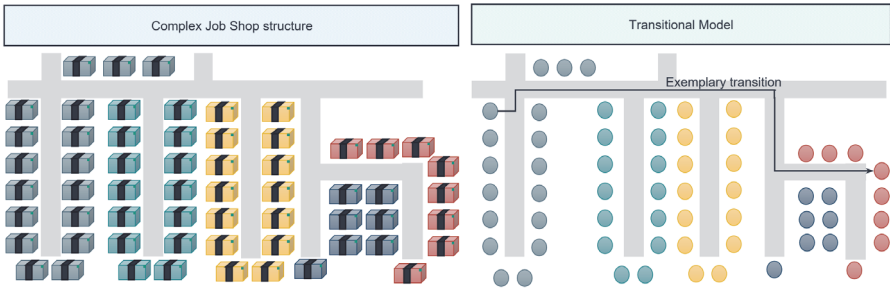

Figure 4.9: Graph based transitional modeling approach as a complex job shop abstraction with an exemplary transition based on May & Maucher, et al. (2021).

Dynamic transitional model

In a complex job shop the complexity originates from the dynamics of the system and the degressive high degrees of freedom. Therefore, not only the graph of transitions that connect equipment is important but also the dynamic material flow on these transitions. Thus, this graph is extended with temporal information about the current jobs on a transition and in a processing equipment as outlined by Figure 4.10. The maximum overlying graph of all transitions that are used more than once is then the sparse transitional model graph $G_{sparse} = (V, E_{sparse})$. The sparse transitional graph is dynamic so that over time novel prior unused transitions can be added. Regarding this graph at a given time has a large

similarity with the knowledge graph based production system model. However, due to the high abstraction important information about the past and current events as well as details are not included. Therefore, the transitional graph can be obtained from a knowledge graph representation easily but not vice versa.

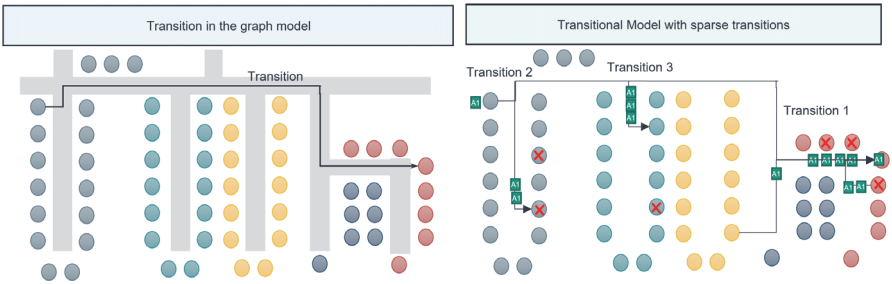

Figure 4.10: Sparse transitional model graph and the temporal graph including material flow represented through boxes on the transition based on May & Maucher, et al. (2021).

The main advantage of this transitional model graph is that it breaks the large convoluted real complex job shop model apart into a set of processing equipment as vertices and a set of transitions or edges connecting these. If each individual transition is regarded and there is significant auto-correlation, a full system view with complexities beyond available computational effort can be avoided. Therefore, the current status of each transition needs to be regarded. The current and past behavior, in other words the start and duration of each product, passing the transition has to be recorded.

Transitional queue construction algorithm

To obtain this transitional information about products that are or have been traveling on one transition from a current knowledge graph based representation or the real production system log data it is necessary to calculate past times and behavior. From this transaction log data of the real complex job shop it is necessary to identify timely sorted transaction pairs which are defined by two consecutive processing operations for each product identifier. Then, the following algorithm introduced in Algorithm 3 can be used to calculate queues over these transitions and in front of machines. It makes use of transition pairs, i.e. a transition that pairs two processing resources, and the past transition data which tracks the transition time and dates of lots traveling on the transitions.

Algorithm 3 calculates the transitional data ordered by actual transition completion time and a queue estimation for any regarded time. To calculate the initial transition and queue data

Algorithm 3 Transitional queue construction algorithm (pseudocode)

```
Require: transition_pairs and past_transition_data
Ensure: Current transition and queue data is available
  initial_transition_data = calc_initial_transition_data(transition_pairs,past_transition_data)
  initial_queue = calc_initial_queue(transition_pairs,past_transition_data)
  for i in transition_pairs do
    lot = i.get_lot()
    transition = (i.get_start_equipment(),i.get_ending_equipment())
    for j in get_lots(transition) do
      if j.arrival_time ≥ lot.arrival_time then
        transition.insert_lot_at_position(lot,j-1)
        break
      end if
    end for
    equip = i.get_start_equipment()
    current_queue = initial_queue.get_equipment(equip).get_current_queue()
    if lot ∉ current_queue then
      current_queue.append(lot)
    end if
    current_queue.get_entry_by_lot(lot).exit_time = i.start_time
    equip = i.get_end_equipment()
    target_queue = initial_queue.get_equipment(equip).get_current_queue()
    target_queue.append(lot)
    target_queue.get_entry_by_lot(lot).entry_time = i.start_time
  end for
```

information from past time snippets can be included or, due to missing information, zero is assumed. This assumption is sufficient as the transitional model requires far less information than other models, so that a long history can be included. The warm start phase, until the initial assumption is not influential anymore, can then be excluded. Based on this initial transition and queue data transitional pairs are iterated to identify all products in transit which are then used to complement the current queue information.

4.3 Intelligent production control architecture for time-constraint adherence

Based on the two introduced simulation and transitional data based modeling approaches for the system architecture the control architecture is described in this section as outlined in Figure 4.11. Therefore, the conceptional model is described based on process flows in Section 4.3.1. Both the foresighted digital twin based gate keeping control and the transitional model based decision making are introduced in Section 4.3.2. The designed algorithm is implemented during operations which is proposed in Section 4.3.3.

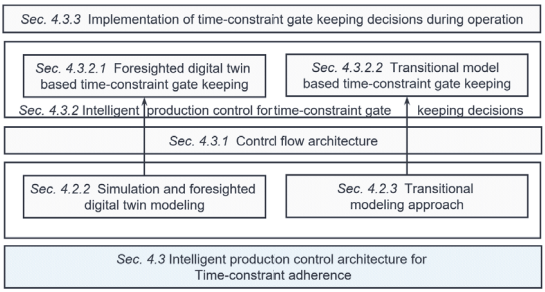

Figure 4.11: Organization of Section 4.3.

4.3.1 Control flow architecture

To minimize the economical and ecological effect of time-constraint violations reducing the number of violations is desirable. As time-constraints are associated with individual products or lots that flow through the production system, multiple production control levers can be taken. As illustrated in Figure 4.12 three production control decision can be controlled to reduce time-constraint violations. First, order release controls the number of novel lots released into the production system. As the production targets should not be compromised through slow order release for lower violation rates, the value of controlling this decision for time-constraint adherence is low. Secondly, order dispatching which controls the transport of lots between resources can be controlled. However, time-constraints are violated to a much larger degree

due to the waiting time than due to pure transporting. Therefore, the focus is on controlling time-constraint gate keeping before any processing starts on the equipment as the third control decision. Thus, those lots with high time-constraint violation probability are held back. The main advantage is that any lot saved from scrapping due to a time-constraint violation is a pareto improvement over the status quo as general system performance is not affected if instead of a high risk time-constrained lot another lot is processed.

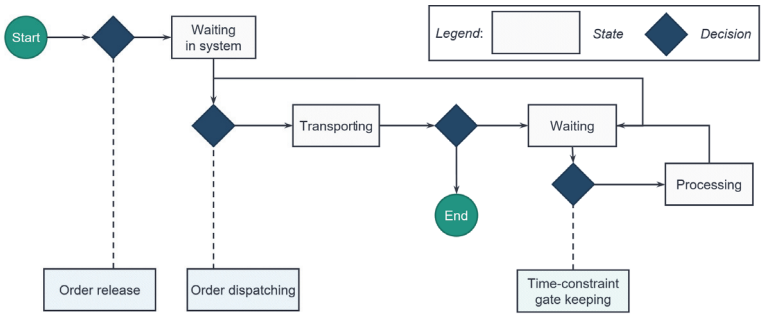

Figure 4.12: Product process flow chart with the regarded PPC decisions and the focus on time-constraint gate keeping decisions.

Gate keeping for time-constraints is a highly dynamic task as illustrated in Figure 4.13. Whenever one lot finished processing at an equipment, here it is lot Z, the next lot or the next batch of lots for batch processing needs to be selected. Within this set of selected lots there should not be any time-constrained lot with a violation probability greater or equal to a threshold. Therefore, the gate keeping decision model restricts the scheduling and dispatching degree of freedom by withholding any of these lots. This decision has to be taken whenever there is a time-constraint lot in the possibly selectable lots $A, B, C, ...$ As illustrated, typically there are non time-constrained lots that can be selected alternatively. Nevertheless, as time-constraint violation results in scrapping not selecting any lot can be better than scrapping. The next gate keeping decision for a withhold lot happens at another time which results in changes in the system, the derived transitional and digital twin model and therefore ensure lots are not unnecessarily withhold for prolonged periods.

First, the current system state is abstracted into the foresighted digital twin model and the transitional modeling which subsequently update the individual models used during decision making. Both models can coexist at the same time, although for individual decision using either the transitional approach or the foresighted digital twin can be sufficient. Based on these updated models each possibly selectable lot is regarded and evaluated with respect to its expected transition time. In other words the expected time of transitioning to the start of

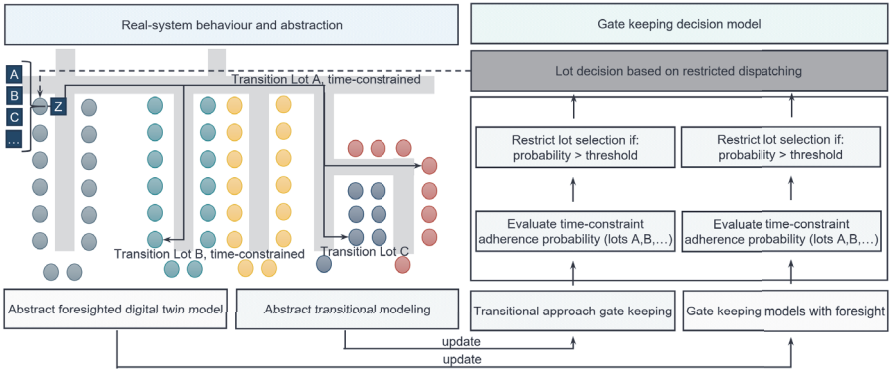

Figure 4.13: Decision framework for time-constrained lots gate keeping decision.

the following process is determined for all time-constraints lots waiting in front of the regarded equipment, i.e. lots $A, B, ...$ in Figure 4.13. Based on these estimates the individual probability of violating the prescribed, known time-limit is calculated. If the probability exceeds a certain threshold the associated lot cannot be selected for processing on the regarded equipment by higher production control levels. As soon as the next equipment can possibly select a time-constrained lot for processing this procedure is repeated. Therefore, quick decision making is required to avoid processing interruptions as the majority of time-constrained lots or lots in general is not critical.

4.3.2 Intelligent production control for time-constraint gate keeping decisions

In order to determine whether or not a particular lot is blocked for processing for the production control the time-constraint gate keeping decision follows Algorithm 4 which requires the time-constrained lot in question, access to the current production system state as well as the selection of risks, in other words coverages, that are tolerated. It can make use of the selected model. Any time-constrained lot can possibly be blocked to avoid time-constraint violations. The selected models are then queried to produce transition time estimates which are consequently evaluated against the time-constraints to identify the violation risk. To preserve computational effort the choice of models can be selected or dynamically controlled, i.e. to include several models for close to risk lots. This algorithms main hyperparameter is the selection of tolerable risk through a threshold. If the models risk estimation exceeds the threshold further processing is prohibited. Due to the sequential nature of this gate keeping decision, any blocked lot will be evaluate in the subsequent gate keeping decisions based on the updated system representation.

Algorithm 4 Time-constrained lot gate keeping algorithm (pseudocode)

Require: time_constrained_lot, production_system_state, risk_threshold, model_choice, release_risk = 1
Ensure: time_constrained_lot is only released if release_risk $\leq$ risk_threshold
  time_constrained_lot.is_blocked = **true**
  foresighted_digital_twin_model.initialize(production_system_state)
    transitional_model.update(production_system_state)
  if transitional_model $\in$ model_choice **then**
    transition = transitional_model.get_transition(time_constrained_lot.current_equipment,
    time_constrained_lot.next_equipment)
    distribution_expected_transitioning_time = transition.calc_distribution(transition.last)
    release_risk = min(1, calc_release_risk(distribution_expected_transitioning_time,
      time_constrained_lot.time_limit))
  end if
  if foresighted_digital_twin_model $\in$ model_choice **then**
    initialize no_of_rollouts, transition_duration_list, no_of_violations = 0
    for $i \in$ range(no_of_rollouts) **do**
      foresighted_digital_twin_model.rollout_until(
        foresighted_time_constrained_lot.last_process
      $\neq$ time_constrained_lot.last_process)
      transition_duration_list.update(resulting_duration(foresighted_digital_twin_model.
        foresighted_time_constrained_lot.transition))
      if transition_duration_list.last $\geq$ time_constrained_lot.time_limit **then**
        no_of_violations = no_of_violations + 1
      end if
      foresighted_digital_twin_model.reset(now)
    end for
    release_risk = max(release_risk, calc_release_risk(transition_duration_list,
      time_constrained_lot.time_limit, no_of_violations, no_of_rollouts))
  end if
  if release_risk $\leq$ risk_threshold **then**
    time_constrained_lot.is_blocked = **false**
  end if

In the following the two distinct approaches for evaluating the risk of this gate keeping decision are presented. Within each approach several models can be implemented and evaluated. This sequential gate keeping decision model avoids traditionally complicated decision making based on the skewed past data. As the goal is to minimize time-constraint violations and which in current production control is done by human supervisors who aim at fulfilling this goal the number of violations is significantly lower than the number of time-constraint adherent transitions. Therefore, end-to-end evaluation of the violations based on the historical data alone is doomed to fail due to heavy bias.

4.3.2.1 Foresighted digital twin based time-constraint gate keeping

The foresighted digital twin is rolled out several times. Each roll out represents the simulation from the instantiated moment until the end of the foresight horizon. While the production control is transferred or mirrored as accurately as possible, each rollout still yields different results. There are various possibilities to recreate the stochastic, dynamic nature with different outcomes in the foresighted digital twin. Hence, this production control follows three steps. First, the rollout mechanism is selected. Second, the simulation of foresight is performed and, third the risk has to be evaluated which is interlinked to the selected rollout mechanism.

Rollout mechanisms

A rollout describes the process of simulating the current simulation status until the end of the regarded period as introduced in Figure 4.14. As an event discrete simulation is a deterministic tool variability is modeled through statistical distributions for process or repairing times, failure and breakdown probabilities etc.. Through controlling this variation and sampling different behavior can be simulated. The following rollout mechanisms have been identified in the course of this work:

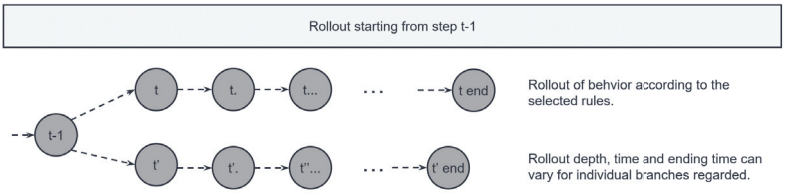

Figure 4.14: General rollout procedure.

1. **Random sampling** describes the simplest rollout mechanism that is identical to a regular DES implementation. Any stochastic values are randomly sampled from the

underlying distribution. The main advantage is that sampling sufficiently often can explore the most likely regions of system development.

2. **Scenario technique** describes an approach in which instead of sampling of stochastic values prior selected values are stored and inserted into the simulation. The benefit lies in the ability to manually inflict desired behavior or circumstances. For instance, a planned sudden change can be manually described to avoid out of distribution sampling when compared to the original distribution.

3. **Seeded random sampling** describes random sampling with a selected seed given to all distributional sampling. Note that each individual distribution needs to be seeded so that behavioral change does not change the sampled values. Seeding has the advantage of creating reproducibility in case of same seeds.

4. **Targeted sampling** describes sampling of values from a subset of the underlying distribution. For instance, the distribution can be cutoff to exclude particularly low or high values. Alternatively, the space can be restricted through the definition of upper and lower quantiles in which sampling is applied. The main advantage of restricting sampling is the possibility to regard worst-case or x quantiled worst-case scenarios.

5. **Any combination of the above** can be used to control the rollout and its value creation for the foresight.

The selection of the rollout mechanism is crucial as the results are heavily influenced and, thus, the rollout results have to be interpreted based on the knowledge of the used and parametrized rollouts. Besides the rollout mechanism and its parameters for one individual rollout, the overall rollout strategy remains paramount. Rollout strategy herein refers to the combination of rollout mechanisms and their parametrization to guide to exploration of the system behavior through the large space of possible futures. Last but not least, the computational effort of the rollout mechanisms as well as the scalability potential through parallelization has to be regarded.

Simulation and rollout during foresight period

Each rollout runs until the end of the pre-selected foresight period in other words until the time-constrained lot has started processing on the target equipment. Parallelization reduces the computational time required to run multiple simulations and enables scalability across different hardware. As shown in Figure 4.15 the rollout strategy, which determines the individual rollout mechanisms and their parametrization during the simulation run, is the basis for controlling the rollout. This leads to different production control decisions being implemented, different behavior and different associated times from the beginning of the rollout and in parallel. In

turn, different sampled time for processing, maintenance, repair etc. lead to a divergence of the individual simulation runs. Parallelization is possible with more than the two depicted mechanisms and parametrizations as suggested in Figure 4.15.

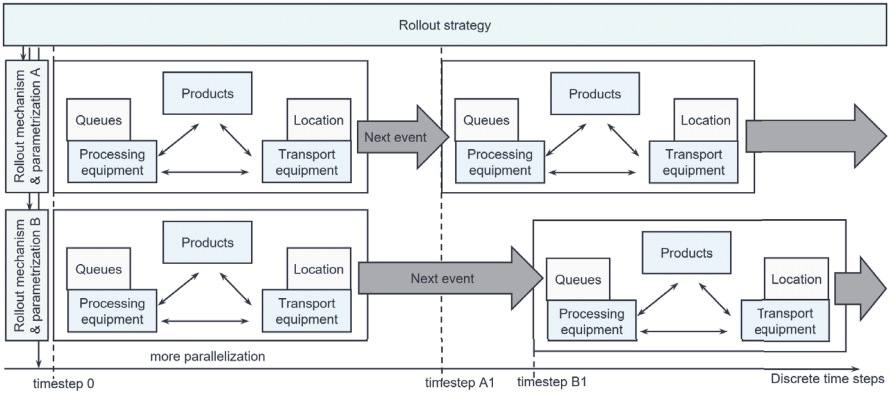

Figure 4.15: Parallelization of rollout behavior during a foresight period.

Risk evaluation

In principle two risk evaluations are possible. First, the algorithmic evaluation based on the results obtained from the different foresight rollouts and the rollout mechanisms and parameters. Second, the evaluation based on priority rules based on the rollout results. The latter can be combined with manual human intervention as an interpretation. Independently of the rollout strategy and knowledge about it, simple rules lead to a risk evaluation. However, the main disadvantage lies in low explainability and comparability of this priority based risk evaluation as parameters, rollout mechanism, strategy and priority rules heavily influence the result.

Thus, algorithmic evaluation is desirable for the application in a complex job shop with a large number of gate keeping decisions during short time intervals. Due to the folding of several statistical distributions during the simulation and rollout process, tampering with the sampling or underlying distribution in the rollout process can heavily influence its results. Therefore, algorithmic evaluation requires random sampling. The presence of seeds however does not prohibit algorithmic evaluations. The simplest algorithmic evaluation is regarding the mean of time-constraint adherence over the different rollouts, i.e. $\frac{n_{adherence}}{n_{adherence}+n_{violation}}$. Based on the law of great numbers for a sufficiently large number of rollouts this value can be taken as the representative adherence probability within the modeled system. However, only few samples

can be taken due to the high computational effort. Thus, alternatively, the recorded transition times can be used to fit a transition time probability distribution. Comparing the allowed risk is possible as the probability of a transition time greater or equal than the time-limit can be calculated based on this distribution. Nevertheless, this distribution is not necessarily close to the ground truth and the distribution type may be unknown.

All in all, the application of the foresighted digital twin is most beneficial if restricted to few cases with high added value through the digital twin to preserve computational effort. Combining a simple random sampling rollout mechanism with a simple algorithmic risk evaluation is most meaningful to avoid behavior far from the real-system and reduce computational effort.

4.3.2.2 Transitional model based time-constraint gate keeping

In fact, the transitional model regards each individual transition separately to determine the time-constraint adherence of a lot that is supposed to be processed and the flow over this transition. May & Maucher, et al. (2021) note the inherent auto-correlation of the transition times regarding an individual transition. Figure 4.16 presents the lowest auto-correlation transition T1. Additionally, T2 is presented as another transition with a high degree of auto-correlation between transition times, as identifiable by the auto-correlation over the time lag t in other words the auto-correlation of a transition time compared to its t-th preceding transition time. Find additional traditional data analysis in the appendix. As past transition times can be retrieved from the transitional model, this inherent auto-correlation can be exploited to predict future transition times.

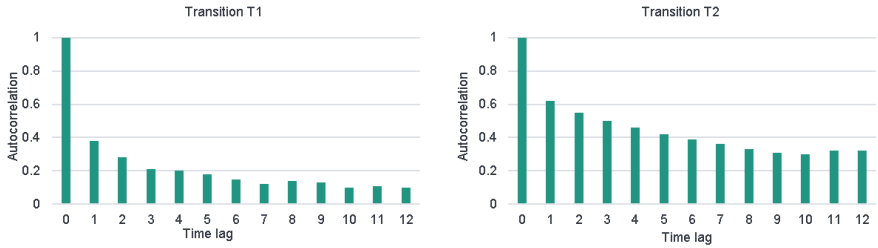

Figure 4.16: Exemplary auto-correlation between consecutive transition times on the same transition based on May & Maucher, et al. (2021).

In order to make use of auto-correlation between values over time, a time-series analysis can be used. As introduced in Section 2.3.5 uni-variate and multi-variate time-series can be distinguished. Fitting a time-series model to the regarded time-series yields a predictor that can be used to predict future transition times as illustrated in Figure 4.17. Typically a

point estimator gives a single value for the predicted transition time (Sadeghi et al. 2015). Comparing this individual value with the known upper time-limit however does not suffice as the uncertainty of this prediction is not regarded and, thus, cannot be taken into account when evaluating the time-constraint adherence probability. Therefore, including this uncertainty into the decision making process yields better results (May & Behnen, et al. 2021). By including this prediction uncertainty from the predictor model, the time-series itself and the uncertainty inherent to the model, the probability density function can be estimated and prediction interval constructed. Based on the confidence of the upper limit prediction interval and the known time-limit it can be evaluated to decide whether or not the violation probability is acceptably small.

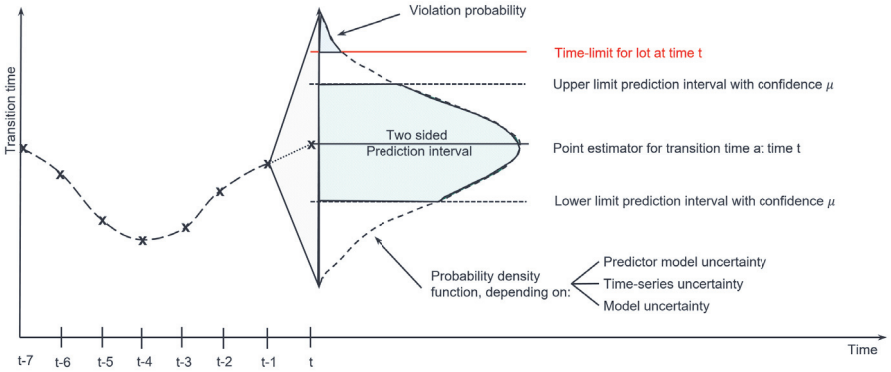

Figure 4.17: Transitional model based on May & Maucher, et al. (2021) and May & Behnen, et al. (2021).

Time series point estimator

As time series stretch associated values over a similar domain, in this case indexed by time steps, only individual realizations are regarded. Therefore, additional information such as the actual time that passed between two time-steps is lost. Nevertheless, time series with significant auto-correlation, in other words with an auto-correlation of about 0.4 or above to the predecessor, are effectively usable to create good prediction models (Farahani et al. 2023). Uni-variate models solely regard past realizations of the time-series to predict the next realization. Alternatively, multi-variate models can be augmented with additional relevant data to improve the prediction. As increasingly high-variate prediction models drive the model complexity and risk for potential overfitting, low multi-variate or uni-variate models are preferable according to Occam's razor if similar performance is obtained (Russell & Norvig

2021). Time series models are fit to decrease an error function described over the predicted and actual value so that the result are good estimations of the next realization.

Time-constraint adherence based on prediction interval

To evaluate the time-constraint adherence based on the prediction model the uncertainty informed prediction interval can be used. The prediction interval with a selected coverage, in other words the confidence in the interval, can be used to obtain the probability of the transition time being smaller or equal to the prescribed time-limit. There are multiple ways and factors to consider when obtaining the prediction interval based on the time-series prediction model. The following uncertainties can be included or interchangeably used:

1. **Predictor model uncertainty** stems from the uncertainty within the prediction model. The prediction model is fit to the visible past of the time-series through a set of parameters that is fixed. As a perfect fit is impossible to obtain these parameters of the model combined with the chosen model lead to an uncertainty within the model compared to the ground truth. If possible this uncertainty should be regarded when constructing the prediction interval to obtain the adherence probability.

2. **Time-series inherent uncertainty** describes the uncertainty within the time-series itself. During the construction of the prediction interval, a suitable basis for the time-series uncertainty has to be selected. Simply put, one can assume a normal distribution or a transformation of a normal distribution to represent the underlying data generation process. With this selection the prediction interval can be constructed around the point estimator.

3. **Model free uncertainty** in form of the Chebyshev model for quantiles (Jørgensen & Sjoeberg 2003) can be selected to create the prediction interval alternatively. If no other assumption on the data generation process can be made Chebyshev still provides a possible solution. However, the interval width of this Chebyshev base model surpasses all other models.

4.3.3 Implementation of time-constraint gate keeping decisions in operations

During operation of the complex job shop, real-time real world behavior drastically changes both the current state of the production system and the transition models. Instant foresighted digital twin instantiation is necessary, so that in parallel to operations a knowledge graph based model and real-time data has to be updated. Due to the high computational effort, the foresight rollout can also be triggered in advance, however, that reduces the accuracy as all events until the actual decision are not regarded. This trade-off is not present in the

transitional model as the time-series and their uncertainty evaluation are comparably fast once the prediction model is trained. However, for the transitional model time series models for all used transitions must be stored. In a large complex job shop with n machines this can lead up to $n(n-1)$ transition models. Nevertheless, the actual number of sufficiently often used transitions is much less. Note that the transition model approach excludes single use transitions or transitions with a long time between each realization as no model can be built or the auto-correlation is not endorsed by any causality.

Thus, the two approaches, namely foresighted digital twin and transitional model, are both kept up to date at the same time. The transitional model can the be used to quickly generate predictions and evaluate decisions and whenever an application is not permissible, i.e. too few data or too complex time-constraints, the foresighted digital twin is used.

4.4 Transition time and adherence prediction

Predicting the transition time and the time-constraint adherence probability is based on both the foresighted digital twin approach as explained in Section 4.4.1 and the transitional model outlined in Section 4.4.2. Each model can perform a point estimation of the expected transition time until the end of the time constraint and an uncertainty informed adherence probability estimation. The models' characteristics are widely differing. While the evaluation of a time series model as in the transitional model requires saving the state of many models at a time but is performed computationally fast, the instantiation of a foresighted digital twin has a high computational effort but only requires one up-to-date knowledge graph as a system representation. Therefore, in the following both approaches are presented and individually regarded as visualized in Figure 4.18. For a later implementation including the near real-time requirements the computational speed has to optimize for the operational performance.

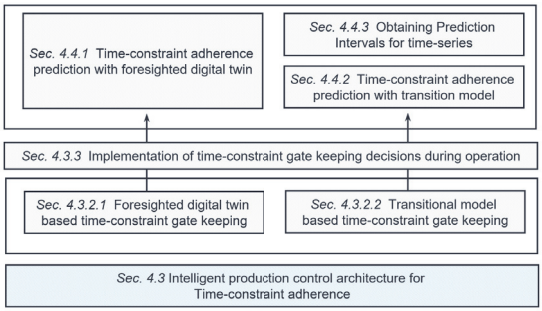

Figure 4.18: Organization of Section 4.4.

In order to make use of both advantages and disadvantages, the precision process is then split according to expected benefit from more computationally intense models over the fast, transition models. As shown in Figure 4.19 at first simple time-constraints that limit two-consecutive processes are regarded with the transitional model. Only if there is no clear decision with the prescribed confidence level about a violation the foresighted digital twin model is used to obtain a clearer result. For time-link areas that restrict the time between two non-consecutive processes or consist of two directly succeeding simple time-constraints the same decision logic is used. However, instead of obtaining one individual prediction from the transitional model, multiple transition times and expected processing times have to be combined and evaluated. Complex time-constraints with overlapping simple time-constraints are evaluated with the foresighted digital twin.

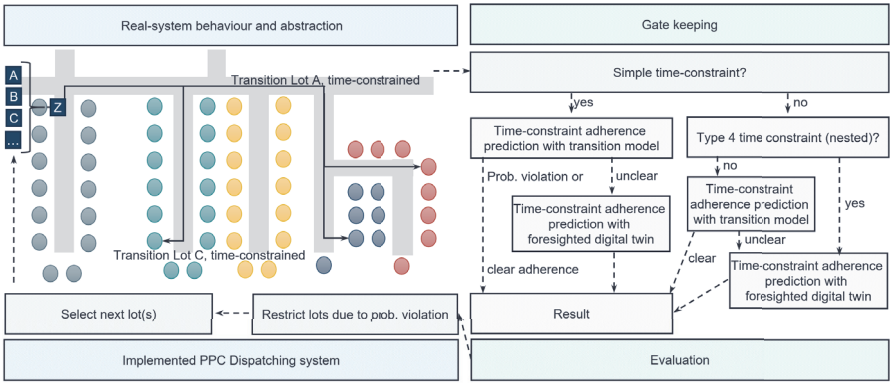

Figure 4.19: General decision process for gate keeping with time-constraint adherence prediction.

4.4.1 Time-constraint adherence prediction with foresighted digital twin

The time-constraint adherence prediction based on the foresighted digital twin makes use of an abstracted digital twin model that is up to date at the decision time as illustrated in Figure 4.20. Within this each individual time-constrained lot is considered separately as the transitions do not necessarily have to be identical and the gate keeping decision restricts only an individual lot. For batching operations different batch combinations with or without one or several time-constrained lots are regarded. For each of these gate keeping decisions several parallel foresight periods are simulated in the digital twin. Therefore, the underlying knowledge graph is copied and instantiated multiple times during the rollout strategy. In the illustration in Figure 4.20 the time-constrained lots and multiple rollouts each are illustrated. Once the time-constrained lot has passed the subsequently required processing operation

these rollouts are halted. Each instantiation is evaluated separately whether or not the time-constraint was violated and regarding the transition time used. As the rollout strategy has to be set before and as discussed earlier a seeded random sampling is used. Based on these individual values and knowledge about the rollout strategy the final evaluation restricts high risk lots as a withholding gate keeping decision due to probable violations. For instance in Figure 4.20 lot C is restricted due to the gate keeping decision and identified possible violation.

As the risk of restricting a lot falsely is associated with low costs or no costs at a time the decision to restrict high risk lots, even if they still have a considerable chance of adhering to the time-constraint, is much more beneficial than falsely letting a violating lot pass. Thus, to speed up the number of rollouts and their evaluation lots are restricted if more than 5% of the regarded instantiations lead to a violation. This can be referred to as a 95% safety interval. As later shown in the benchmarking and implementation this parameter can be varied, however, it should reflect the complexity in a complex job shop so that higher risks break the promise of significantly reducing time-constraint violations.

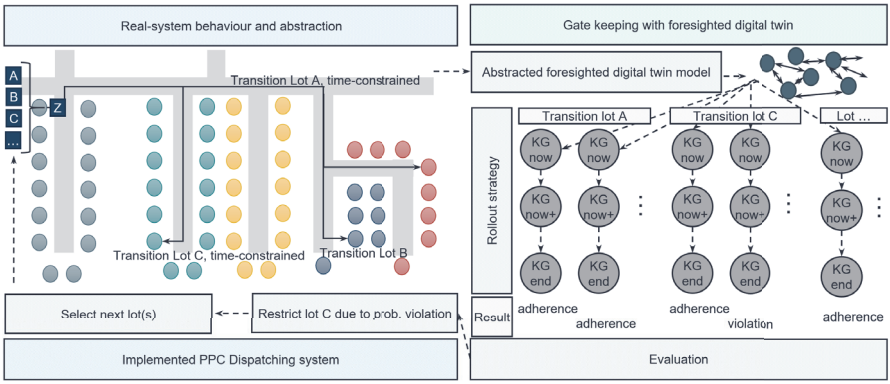

Figure 4.20: Using the foresighted digital twin to predict time-constraint adherence for the gate keeping decision.

Regarding the final objective of reducing time-constraint violations through intelligent production control a higher safety interval corresponds to a lower violation rate. However, if no time-constrained lots are released at all no violations go hand in hand with clogging the production system as the number of lots in a system is not infinite. Thus, the aim is to minimize time-constraint violations without heavily impeding the production flow and throughput which translates into a safety interval smaller than 100%. In fact the safety interval and rollout strategy constitute the set of hyperparameters of the foresighted digital twin based model.

These parameters can be tuned based on historic data for each individual use-case by ex post evaluation of the desired trade-off obtained.

Complex time-constraints and timelink area constraints

In order to predict the adherence or violation of complex time-constraints, i.e. these time-constraints with more than one single transition with a time-limit, the uncertainty of the prediction increases as longer time-spans have to be simulated. Thus, the evaluation and hyperparameters can be changed. Nevertheless, the overall decision process remains identical to simple time-constraints as the foresight obtained through several simulations from the status quo forms the basis.

A longer time-horizon increases the windows of opportunity to change dispatching and sequencing as a result from standard changes during manufacturing in a complex job shop. Theoretically, known changes can also be coupled to the foresighted digital twin. However, in this scenario of time-constraint violations such changes are unknown in advance and therefore not regarded.

Gate keeping decisions during foresight period

While general production control can be transferred into the foresighted digital twin, the gate keeping decision itself is hardly transferable if the decision is based on a simulation rollout. That would lead to a combinatorial explosion as for each time-constrained lot that can be selected during the foresight period, another foresighted digital twin or several had to be started. Likewise, the simulation would take exponentially longer with longer simulation times. While the gate keeping decision has a large influence on the individual lot and its transition time, the overall influence on another lot rarely turns into a large effect. As any lots arriving at the target equipment after the time-constrained lot are in the queue behind them, random or seeded random sampling is used for the gate keeping decisions of other lots during a simulation process.

4.4.2 Time-constraint adherence prediction with transition model

The transitional model consists of one or more prediction models for each sufficiently used transition. With each of these models, the goal is to evaluate the time-constraint adherence probability so that the gate keeping decision can withhold this lot if necessary. Thus, instead of only evaluating the transition time point estimator $\hat{y}$ the upper limit of the prediction interval has to be compared to the prescribed time-limit. Doing so captures the uncertainty in the prediction and time-series. The confidence of the next time series value being lower than this one-sided prediction interval is described with α which corresponds to a probability of exceeding this upper limit of $1 - \alpha$. Based on the student's t-distribution this confidence can

be converted into the factored standard deviation depending on the variance $Var(e)$. Thus, formally, the upper prediction interval limit can be described with Equation 4.1, as visualized in Figure 4.17.

$$\hat{y} + t_{1-\alpha} \times \sqrt{Var(e)} \tag{4.1}$$

This upper prediction interval, described as the value that with a $1 - \alpha$ confidence will not be exceeded by the next time series value, has to be compared to the upper time limit. As discussed only upper time limit time-constraints are regarded. The transitional approach for an individual prediction is for now limited to simple time-constraints that fully lie on the same transition so that the next time-series realization can be evaluated as the transition time corresponding to the time-constrained lot in question. Hence, the upper prediction interval should be smaller or equal to the prescribed time limit d^u as shown in Equation 4.2.

$$\hat{y} + t_{1-\alpha} \times \sqrt{Var(e)} \leq d^u \tag{4.2}$$

Equation 4.2 can be restructured to obtain the time-constraint adherence probability by computation, as follows:

$$t_{1-\alpha} \leq \frac{d^u - \hat{y}}{\sqrt{Var(e)}} \tag{4.3}$$

By using Equation 4.3 the corresponding probability value can be determined from the cumulative density function of the student's t-distribution. Comparing the obtained probability with the selected confidence level $1-\alpha$ the violation of this time-constraint is likely if the probability is lower than the confidence interval. Thus, Equation 4.3 illustrates the transitional modeling approach as it is interpretable as the time-constraint adherence probability. Therefore, each transition model consists of first a regression model used to predict the point estimate for the next transition time based on the auto-correlation of transition times as the underlying approach introduced before. Secondly, the model consists of the corresponding prediction interval to obtain the probability of time-constraint violations.

Thus, the known time limit, the observation, in other words the ex post realization of the transition time and the known upper limit of the prediction interval can be regarded as illustrated in Figure 4.21. There are three interesting cases starting with the actual ex post transition time (the observation) exceeding the upper prediction interval and time limit. This is a wrongly unidentified time-constraint violation (false positive) and should be avoided.

Through arbitrarily low α this case can be avoided. However, this comes at high operational expenses as the number of withhold lots increases accordingly. Thus, the confidence should be kept as low as possible. In turn this has two more interesting cases. Secondly, in a false alarm, the upper prediction interval exceeds the time limit and the gate keeping decision restricts the lot. This is not optimal for the operational process and should be minimized with a higher α but in general the costs associated are significantly less than for violations (May & Maucher, et al. 2021). Thirdly, if the realized value actually exceeds the time limit but was identified before a time-constraint violation could be inhibited. The latter case is only observable in ex post comparisons as outlined in the benchmarking and results. Last but not least, in the standard case in which the upper prediction interval and the realized transition time does not exceed the time limit is most common.

Figure 4.21: Transitional model based gate keeping based on one-sided prediction interval.

The main advantage of this approach is that each transition is regarded and evaluated separately. Therefore, the peculiarities on each transition, such as batching or not batching, are implicitly included in the observed past transition times. Moreover, each lot has its own time limit for the time-constraint, so that the upper limit that should not be exceeded in fact varies, in contrast to the simplified illustration in Figure 4.21. Nevertheless, the observed transition time can be used for the time series prediction, irrespective of the actual time limit and irrespective of the skewed data set with a much larger share of time-constraint adherences than violations in a real-world complex job shop. Additionally, any transitioning of a lot on this transition, whether or not time-constrained, leaves a trace and observation that can be used to improve and update the model. Last but not least, lots that were shifted in a transition due to breakdowns and machine dedications are also part of the model, and hence both standard and more complex cases are included. Thus, the number of lots one can learn from increases by a large factor, as only a number of lots at each transitions is time-constrained depending on the technological requirements as outlined in Section 2.1.

This can be exploited to reduce the uncertainty and, hence, the prediction interval width. In turn this improves the transitional models prediction capabilities and operational performance due to more accurate gate keeping. Using one-sided prediction intervals is rarely regarded and to understand the theoretical background and the derived formulas the following section introduces the prediction interval approach used in this thesis.

4.4.3 Obtaining Prediction Intervals for time-series

In the following the process to obtain prediction intervals and value their quality with dedicated loss functions is introduced. Firstly, prediction intervals are formally described and secondly their evaluation is presented. Prediction intervals for sophisticated machine learning models can be obtained with methods introduced in the section hereafter.

4.4.3.1 Introduction to Prediction Intervals

Estimations from a statistical point of view are split into point estimations and interval estimations. Point estimations provide a probability or a single value while interval estimations derive a range with an associated degree of uncertainty (Handl & Kuhlenkasper 2018). Confidence intervals are a prime example for interval estimations as they give the range of estimates of an unknown parameter, such as the mean or variance, according to a selected confidence level. In light of time series and general predictions as introduced in paragraph 2.3.5 prediction intervals regard the uncertainty associated with the single next data point. Concretely, a prediction interval for the next data point estimates the interval from which it is sampled. Consider the independently and according to a normal distribution $N(\mu, \sigma^2)$ identically distributed stochastic variables $X_1, ..., X_n$ where the actual μ and σ are unknown. Using the student's t distribution and estimators $\bar{X}$ and s the corresponding two-sided confidence and prediction interval can be obtained. Selecting a confidence, and similarly a coverage in the prediction interval, of $1 - \alpha$, the confidence interval can be be given as Equation 4.4 while the prediction interval can be given as Equation 4.5.

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{4.4}$$

$$\left[\bar{X} - t_{n-1;1-\alpha/2}s\sqrt{1+\frac{1}{n}},\ \bar{X} + t_{n-1;1-\alpha/2}s\sqrt{1+\frac{1}{n}}\right] \tag{4.5}$$

Comparing the confidence and prediction interval shows that the prediction interval exhibits a higher uncertainty leading to a larger interval as elaborated in Equation 4.6. While the

confidence interval width converges to zero with increasing sample size n, the prediction interval's width is always larger than zero despite arbitrarily large sample sizes.

$$t_{n-1;1-\alpha/2}s\sqrt{1+\frac{1}{n}} \geq t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}$$
$$\sqrt{1+\frac{1}{n}} \geq \sqrt{\frac{1}{n}} \tag{4.6}$$

As the aim of this thesis is the intelligent control of time-constraints in complex job shops which limit the maximum transition time future realizations of the transition time have to be regarded. As a result, prediction intervals are used throughout. As the limit is an upper limit for the transition time, the prediction interval in fact only needs to be one sided, which can be given in Equation 4.7.

$$\left(-\infty,\ \bar{X} + t_{n-1;1-\alpha}s\sqrt{1+\frac{1}{n}}\right] \tag{4.7}$$

4.4.3.2 Loss function derivation for Prediction Intervals

In order to compare several models that aim at generating sufficiently good prediction a model assessement should evaluate their ability to produce well fit prediction intervals. Ultimately, the best fitting model can then be used to control the gate keeping decision for time-constrained lots. While point estimators can easily be compared with standard methods, such as the Mean Absolute Error (MAE) or the Mean Squared Error (MES) which apply an absolute or squared penalty function on the errors, in other words the difference between predicted and observed values. Thus, greater deviations lead to higher errors and ultimately to a worse performance evaluation. The simplest conversion of this approach to prediction intervals lies in the trivial calculation of a classification loss on the number of observations that were realized in the predicted interval. With the free selection of α any coverage $1-\alpha$ can be realized, so that broad coverages that include all data realization in the prediction interval are favored. Such infinitely large prediction intervals, however, provide no meaningful model. Thus, the achieved coverage and the interval width need a trade-off to produce suitable prediction intervals. Therefore, prediction interval loss functions that suit the actual application need to be derived.

Desirable properties for interval loss functions

In other cases loss function derivation for prediction intervals has already been studied. Askanazi et al. (2018) identify five properties desirable or any prediction interval loss function L to be created. First, an inverse trade-off between coverage and length is considered in L. To prevent arbitrarily large prediction intervals, as explained above, this property is

crucial. Second, the Casella paradox stating that Loss functions that too strongly enforce short intervals widths creates stark interval miscalibration (Casella et al. 1993) should be avoided. Third, to ensure consistency the shortest well-calibrated prediction interval has to minimize the loss. For instance, if the underlying Data Generation Process (DGP) is known and the shortest well-calibrated prediction interval is obtained there cannot exist another wider prediction interval forecast at the same coverage $1-\alpha$ which achieves a lower loss L. Forth, the Loss function evaluation of prediction intervals must be obtainable without knowledge about the DGP to ensure applicability. Fifth, the forecasting quantiles to obtain the prediction interval shall not be necessary to calculate the Loss L and finally evaluate the prediction intervals.

Description of the Winkler Loss

The Winkler Loss $L_{winkler}(y, d, \lambda)$ constitutes the most frequently used loss function for prediction intervals (Askanazi et al. 2018). Thus, the following paragraphs describes the Winkler loss and evaluates the applicability to the prediction intervals that shall be used to predict time-constraint adherence. The Winkler Loss can be constrained to the interval $[0, \infty)$ where a lower loss indicates a better prediction interval. The following notation is used: d denotes the length of the interval and d^u as well as d^l denote the respective upper and lowever prediction interval bounds. Then the Winkler Loss can be given as in Equation 4.8. $|d|$ constitutes the first term that penalizes large interval widths. The second term $\lambda_l(d^l - X)1\{X < d^l\}$ penalizes values lower than the lower bound, whereas the third term $\lambda_u(X - d^u)1\{X > d^u\}$ penalizes outliers larger than the upper bound. To find a balance between the interval width and coverage the parameters λ_l and λ_u are used and should correspond to $1/\lambda_l + 1/\lambda_u = \alpha$ (Gneiting & Raftery 2007).

$$L_{winkler}(X, d, \lambda) = |d| + \lambda_l(d^l - X)1\{X < d^l\} + \lambda_u(X - d^u)1\{X > d^u\} \tag{4.8}$$

Reasons for using the Winkler Loss

When measuring the Winkler Loss with the five desired properties set forth by Askanazi et al. (2018) it becomes apparent that only the first and second property are fulfilled. The Winkler Loss evaluation is based on the quantiles that were used to obtain the interval. As both α and λ depend on the selected quantile the Winkler loss does not fulfill the desired properties four and five. As a matter of fact this can be used to show that the Winkler Loss also prefers intervals that are closer to the selected quantiles compared to those further away which leads to a violation of the third desired property (May & Maucher, et al. 2021). Since the later obtained models that are evaluated with the Winkler loss are quantile based and not derived from nonparametric models, these violations can be accepted (Gneiting & Raftery

2007). After all the Winkler is good at given a desired coverage finding the best prediction interval. This is particularly important for the desired application as not only the best interval but the best interval in conjunction with a selected coverage that should be achieved to reach business purposes is required.

Derivation of a custom Winkler Loss

Another major argument for using the Winkler loss lies in the possibility to adjust it to create an individualized loss function. As time-constraints violations come from maximum time limits for transition times that are exceed only a one-sided prediction interval shall be considered, namely the upper prediction interval. Thus, the one-sided prediction interval Winkler loss has to derived. The two-sided interval score uses the prediction of multiple quantiles $r_1, \ ..., \ r_k$ to derive the Winkler loss as shown in Equation 4.9 based on Gneiting & Raftery (2007). Concretely, $k = 2$, l^u and l^l represent the upper and lower bounds corresponding to the $1 - \frac{\alpha}{2}$ quantile for the two-sided interval score. Following the suggestion from Gneiting & Raftery (2007) α_1 is therefore set to $\frac{\alpha}{2}$ while α_2 is set to $1-\frac{\alpha}{2}$ and s and h are set to $s_1(x) = s_2(x) = 2\frac{x}{a}$ and $h(x) = -2\frac{x}{\alpha}$ as both function require to be polynomial in x. While h can be in general arbitrary, s must be non-decreasing. Equation 4.10 this two-sided Winkler loss where the scoring rule sign is reversed for obtaining the negatively oriented interval score.

$$L(r_1, ..., r_k; X; \alpha) = \sum_{i=1}^{k} [\alpha_i s_i(r_i) + (s_i(X) - s_i(r_i))\mathbb{1}\{X \leq r_i\}] + h(X) \tag{4.9}$$

$$L_{two-sided}(d^l, d^u; X; \alpha) = (d^u - d^l) + \frac{2}{\alpha}(d^l - X)\mathbb{1}\{X < d^l\} + \frac{2}{\alpha}(X - d^u)\mathbb{1}\{X > d^u\} \tag{4.10}$$

As time-constraint adherence in gate keeping decisions for intelligent production control focuses exclusively on the upper limit of the prediction interval the one-sided interval score must be obtained. Thus, $k = 1$ is selected as solely the upper limit d^u shall be regarded. Additionally, $\alpha_1 = 1 - \alpha$ is selected as the $1 - \alpha$ coverage level's upper bound is considered. Accordingly, $s_1(x) = \frac{x}{\alpha}$ and $h(x) = -\frac{x}{\alpha}$ are selected. Following some rearrangements the inserted formulas yield the once again reversed scoring rule to finally constitute the one-sided Winkler loss function as shown in Equation 4.11 (May & Maucher, et al. 2021) which is analogously developed to the two-sided Winkler loss as shown in Gneiting & Raftery (2007).

$$\begin{aligned} L_{one-sided}(d^u; X; \alpha) &= (1-\alpha)\frac{d^u}{\alpha} + (\frac{X}{\alpha} - \frac{d^u}{\alpha})\mathbb{1}\{X \leq d^u\} - \frac{X}{\alpha} \\ &= \frac{d^u}{\alpha} - \frac{X}{\alpha} - d^u + \frac{1}{\alpha}(X - d^u)\mathbb{1}\{X \leq d^u\} \\ &= -d^u + \frac{1}{\alpha}(d^u - X) + \frac{1}{\alpha}(X - d^u)\mathbb{1}\{X \leq d^u\} \\ &= -d^u + \frac{1}{\alpha}(d^u - X)(1 - \mathbb{1}\{X \leq d^u\}) \\ &= -d^u + \frac{1}{\alpha}(d^u - X)\mathbb{1}\{X > d^u\} \end{aligned} \tag{4.11}$$

$$L_{one-sided}(d^u; X; \alpha) = d^u + \frac{1}{\alpha}(X - d^u)\mathbb{1}\{X > d^u\} \tag{4.12}$$

Beyond the derivation of the one-sided Winkler loss it becomes apparent that the approach has to be adapted to evaluate not only one data point as shown in Gneiting & Raftery (2007) but to regard a larger data set. Thus, the overall loss is extended as an average of multiple data points. By doing so models of different complexities become comparable on a larger scale. The advantage of a non scaled average is that additionally compound models that use different prediction intervals for different subsets can still be effectively compared with one another. As a result the finally used custom one-sided loss function used can be given as Equation 4.13.

$$L_{one-sided}(d^u; X_i, ..., X_n; \alpha) = \frac{1}{n}\left(\sum_{j=1}^{n} d_i^u + \frac{1}{\alpha}\sum_{j=1}^{n}[(X_j - d_j^u)\mathbb{1}\{X_j > d_j^u\}]\right) \tag{4.13}$$

4.4.3.3 Basic prediction interval estimators

To obtain the prediction intervals as outlined before and shown in Figure 4.22, relevant assumptions as to the underlying data generation process have to be taken. First, no assumption leads to the application of Chebyshev to avoid estimation of a probability density function. Alternatively, the historic observations can be regarded as being i.i.d. distributed to fit a distribution which can be used to obtain prediction intervals, for instance a normal or lognormal distribution. As an alternative the predictor inherent uncertainty can be used to generate the prediction interval. For the basic interval estimation predictor inherent uncertainty is not yet regarded.

Regarding historic transition times can support identifying suitable distributions of those and the required underlying information that can help in creating the upper prediction interval.

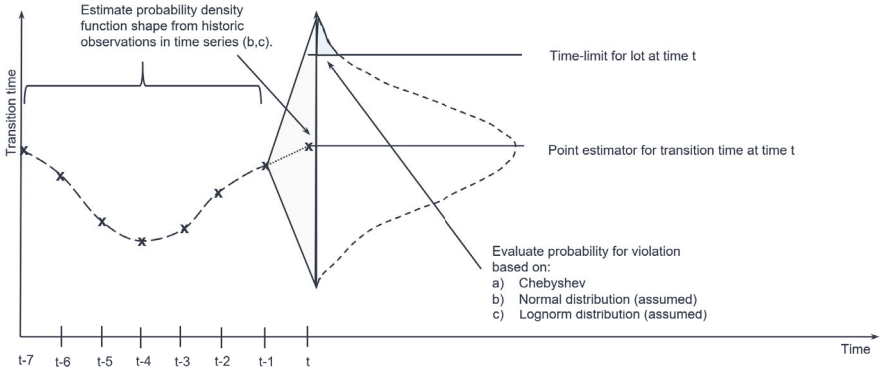

Figure 4.22: Three approaches to obtain basic prediction interval estimators (a) Chebyshev, (b) assumed Normal distribution or (c) assumed Lognormal distribution.

Thus, sampling transitions with a large available number of observations during a predefined observation space can be visualized as in Figure 4.23 on the left side. The transitions are numbered to obscure the actual material flow but the data remains unchanged. A right skewed normal distribution could be used to approximate this behavior and improve from a distribution free setting. Given the underlying multiplicative independent random variable product the central limit theorem suggest a normal distribution in the log domain (Gneiting & Raftery 2007). Nevertheless, the distribution has a very long tail from singular very long transition time cases. Through a natural logarithm transformation large values can be moved closer to the mean to deal with the long tails. Likewise, smaller values would be stretched out to deal with the heavy right-skewness. The results can be seen on the right side of Figure 4.23 and indicate a potential normal distribution after logarithmic transformation.

Testing for normality with the Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test can be used to test if an observed distribution fits an assumed underlying distribution that has to be predefined for the test. Thus, it is a nonparametric goodness-of-fit test (Dodge 2008). H_0 describes the null hypothesis which assumes the both accumulated functions $F(x)$ and $G(x)$, that describe the observed and assumed distributions, are sufficiently equal. T is the statistical test which is the maximum absolute difference between the two distribution functions. Therefore, the test can be described in Equation 4.14 as:

$$T = \sup_x |F(x) - G(x)| \tag{4.14}$$

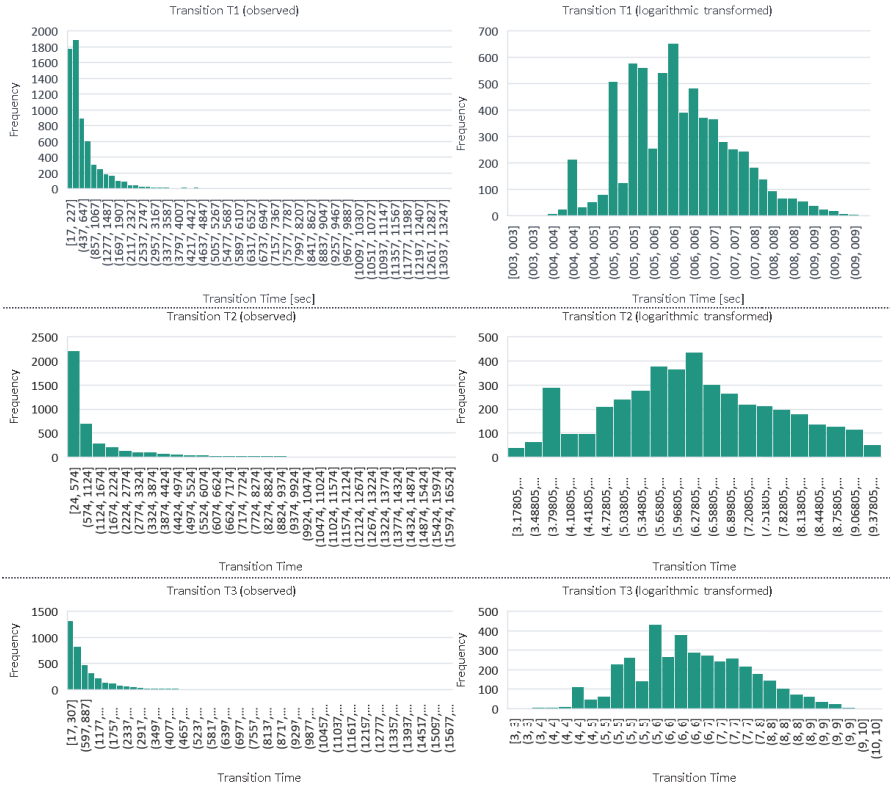

Figure 4.23: Observed transition time distribution and logarithmic transformation.

Exemplarily the Kolmogorov-Smirnov test is reported for the transition time distribution T1 having the largest available data and visually seems to correspond to a normal distribution after transformation. Thus, a rejection of the hypothesis H_0 in this case would likely also correspond to a rejection in the other transitions. The critical value of test T can be reported as 0.0473. The $K_{n;1-\alpha/2}$ quantile for the two-sided Kolmogorov-Smirnov can be derived from respective table. However, for quantiles with large sample size (greater than 35) the approximation Equation 4.15 should be used (May & Maucher, et al. 2021):

$$K_{n;1-\alpha/2} = \frac{\sqrt{ln(\frac{2}{\alpha})}}{\sqrt{2n}} \qquad 4.15$$

This yields for the sample size greater than six thousand and the α level of 0.05 the critical value 0.0165. As $0.0473 > 0.0165$ and hence the statistical test value being greater than the critical value the H_0 hypothesis has to be rejected with a 0.05 confidence level. Thus, the assumption of equal normal distributions cannot be confirmed which, regarding the visual comparison, is surprising.

Taking one step back the usage of the large sample size Kolmogorov-Smirnov approximation might be responsible for this rejection. Filion (2015) confirms this influence of the sample size as the critical value, for large sample sizes, becomes very small. As the observed distributions are rarely ideally following the symmetry requirement this often results in H_0 hypothesis rejections. Thus, only random subsets of the observed period are regarded for the following Kolmogorov-Smirnov test. For instance, Figure 4.24 in addition to the original test reports the test of first 30 records of transition T1. The latter leads to a critical value of 0.2417 which is larger than the statistical test T resulting in 0.1095. Thus, herein the H_0 hypothesis is not rejected. Note that not rejecting is not directly implying that it should be accepted. Therefore, from a purely statistical point of view the logarithmic transformation does not result in a normal distribution of the transition times. Nevertheless, this error might be acceptable as both smaller sample sizes and visual comparisons indicate only minor deviations.

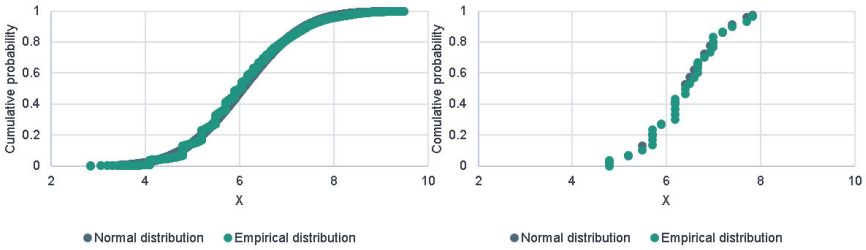

Figure 4.24: Kolmogorov-Smirnov test comparison for large sample size and selection of the first 30 records based on May & Maucher, et al. (2021).

As explained, for an observed set of transition records that form a distribution the upper prediction interval can be given by Equation 4.16 with the mean $\bar{y}$, standard deviation s, sample size n for any given coverage level. Thus, the coverage level needs to be defined in the following. Firstly, most importantly the underlying distribution has to be assumed.

$$\bar{y} + t_{n-1;1-\alpha/2} s \sqrt{1 + \frac{1}{n}} \tag{4.16}$$

Base case model with Chebyshev's inequality, assumed normal distribution and assumed logarithmic transformed normal distribution

There are two main approaches to obtain the prediction intervals. Firstly, the prediction interval model can be based on the point estimator inherent uncertainty which might be quantified with a holdout data-set as introduced in Section 4.4.3.4. Secondly, the underlying distribution of the time series has to be identified and prediction interval quality measured through the previously defined Winkler loss. The latter is introduced in the following with three different approaches assuming no distribution with Chebyshev, assuming a normal distribution or a logarithmic transformed normal distribution. The latter might not be statistically validated through the Kolmogorov-Smirnov test as explained earlier, however, they can still significantly reduce the prediction interval width.

In general, the Chebyshev inequality is known to be applicable for obtaining prediction interval without any knowledge or assumption of the distribution (Jørgensen & Sjoeberg 2003). The assertions deducted with Chebyshev's inequality work without any known probabilities so that boundaries are deducted instead of precise values. Based on Chebyshev's inequality in Equation 4.17 which is two-sided the one sided form, named Chebyshev Cantelli inequality, can derived:

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \tag{4.17}$$

$$P(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \tag{4.18}$$

In short the one-sided Chebyshev Cantelli inequality in Equation 4.18 give the upper bound for realization probability of the stochastic X being above the specified value of a greater then the mean. Thus, $\mu + a$ make up this upper bound and the probability of exceeding it is the prescribed α. They can be associated as follows:

$$P(X - \mu \geq a) = P(X \geq \mu + a) = P(X \geq d^u) \leq \frac{\sigma^2}{\sigma^2 + a^2} = \alpha \tag{4.19}$$

With given α based on Equation 4.20 one can get a based on the standard deviation σ as $a = \sigma\sqrt{\frac{1}{\alpha} - 1}$. In turn the standard deviation can be estimated from the regarded sample and the mean $\bar{y}$ can likewise be obtained. All in all, this can be combined to obtain the upper prediction interval as in :

$$d^u = \mu + a - \mu + \sigma\sqrt{\frac{1}{\alpha} - 1} = \bar{y} + s\sqrt{\frac{1}{\alpha} - 1} \tag{4.20}$$

In a similar vein the prediction intervals can be constructed for the normal distribution and the logarithmic transformation model. All distributions are fit to the observed transitional data, which results in a large number of fits. Then the same approach is used to obtain the prediction intervals. As these assumptions restrict the form of the probability density function the underlying rationale is that tigther and hence more favorable prediction intervals can be obtained.

4.4.3.4 Prediction Intervals for Neural Networks

As an alternative to directly estimating the prediction error based on the overall distribution outlined above one can split the prediction error into the aleatoric uncertainty of general noise and the epistemic uncertainty of the used prediction model. In the following the statistical foundation and underlying approach is presented. Ultimately, this section introduces the approach to obtain prediction intervals for a single prediction from the prediction models uncertainty quantification.

Splitting the total variance of the prediction error

When regarding the prediction error the original observation X_i, in a similar vein to the time series methods introduced in paragraph 2.3.5 can be given as a composition of the signal x_i, i.e. the part that can be learned and stems from an understandable stochastic process, and a noise based error term ϵ_i (Chatfield 2001). For statistical purposes it is often assumed that the noise in the observation is identically, normal distributed around zero with constant variance. Then the observation X_i can be given as shown in Equation 4.21.

$$X_i = x_i + \epsilon_i \tag{4.21}$$

Then using the model to produce a prediction $\hat{X}_i$ and comparing it with the actual observation X_i yields the prediction error e_i given as the difference shown in Equation 4.22.

$$e_i = X_i - \hat{X}_i = [X_i - \hat{X}_i] + \epsilon_i. \tag{4.22}$$

Consequently, Khosravi et al. (2011) give the total variance of the prediction error as Equation 4.23 if both terms on the right hand side of Equation 4.22 are statistically independent. Then the noise's variance σ_ϵ^2 can be understood as the aleatoric, irreducible uncertainty

intrinsically linked to the underlying distribution in the date (Abdar et al. 2021). Following the approach from Zhu & Laptev (2017) this variance of the noise can be estimated through dividing the squared distance between observations X_i and multiple predictors performing an ensemble prediction $\hat{X}_i$ over the values in a test set n_t as shown in Equation 4.24.

$$Var(e) = \sigma_i^2 = \sigma_{\hat{X}_i}^2 + \sigma_{\hat{\epsilon}}^2. \tag{4.23}$$

$$\sigma_{\hat{\epsilon}}^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(X_i - \hat{X}_i\right)^2 \tag{4.24}$$

The model inherent variance of the prediction $\sigma_{\hat{X}_i}^2$ on the contrary is based on misspecified models or insufficiently used information (Khosravi et al. 2011). It is in the context of machine learning usually denoted as the epistemic uncertainty that is a feature of the used model and has to be considered when selecting the models. Unavoidable uncertainty is caused by random noise from within the data, in contrast to the model's prediction uncertainty that can be reduced with quality data and suitable model selection. A thorough derivation of the used quantification appraoch is given in the following.

Obtaining neural network (NN) prediction intervals

Neural networks possess the capability to solve many different tasks, yet their function is still often considered as a black box. Some insight comes from analyzing individual neurons that may be more active for certain results, i.e. for picture classifications a set of neurons is only active when the respective class is triggered, their importance to the performance is not increased compared to other neurons (Morcos et al. 2018). To extend the applicability of neural networks and use or reduce their epistemic uncertainty the underlying uncertainty must first be quantified with uncertainty quantification (UQ). Within the literature two UQ methods stand out due to their applicability to a wide range of neural networks: bootstrapping and monte carlo dropout. The uncertainty quantification finally used is derived from these.

Bootstrapping for quantifying uncertainty of neural networks

Bootstrapping is the most popular method in literature to construct prediction intervals from the quantified uncertainty in neural networks (Khosravi et al. 2011). The main approach is as follows: the distribution of a population is approximated by examination of the distribution of random samples of the population. There are several bootstrapping types that slightly alter the procedure, for instance smooth, parametric, paired and wild (Kabir et al. 2018). Concretely, for neural networks an ensemble that consists of N neural networks has to be trained. Then,

each network's prediction in the ensemble is differentiated as $\hat{X}_{i,n}$. $\hat{X}_i$ is then obtained as the average of the ensemble predictions. The variance associated with these N predictions can then be used as an estimate for the model uncertainty $\sigma^2_{\hat{X}_i}$. The average and variance can be calculated with Equation 4.25 and Equation 4.26 (Khosravi et al. 2011).

$$\hat{X}_i = \frac{1}{N} \sum_{i=1}^{N} \hat{X}_{i,n} \tag{4.25}$$

$$\sigma^2_{\hat{X}_i} = \frac{1}{N-1} \sum_{n=1}^{N} (\hat{X}_{i,n} - \hat{X}_i)^2. \tag{4.26}$$

For individual tasks training an ensemble of multiple neural networks as predictor which is computationally significantly more expensive by approximately the factor N is tolerable to obtain distribution predictions and model uncertainty. With increasing complexity of individual models this becomes hardly possible for larger problems. Given the large number of transitions and hence large number of prediction models that need to be trained the application bootstrapping is not feasible in this use-case.

Monte Carlo dropout for quantifying uncertainty of neural networks

In contrast to bootstrapping, monte carlo dropout obtains the ensemble predictors from different dropouts on the same neural network. Originally, dropout is technique widely used for regularization to prevent overfitting in deep neural networks. During every training step a random selection of neurons and their associated connections and weights is removed for only the individual training step. As a result the predictor is comparably agnostic to excluding a small random selection of neurons. Figure 4.25 presents a comparison of a neural network with droupout applied during this training step and a general fully conected neural network.

Given a neural network trained with dropout applied, the model uncertainty can be determined through a prediction with different random neuron sets in the dropout. In general, it can be shown that the application of dropout in front of each hidden layer is comparable to approximating a probabilitic deep Gaussion process (Gal & Ghahramani 2016). Concretely, $\hat{X}_{i,n}$ predictions can be obtained with N stochastic forward passes. Then, Zhu & Laptev (2017) use the obtained sample variance to estimate the model uncertainty $\sigma^2_{\hat{X}_i}$. Then,Equation 4.27 gives the model uncertainty obtained by applying dropout N times for the final prediction.

$$\sigma^2_{\hat{X}_i} = \frac{1}{N-1} \sum_{n=1}^{N} (\hat{X}_{i,n} - \hat{X}_i)^2. \tag{4.27}$$

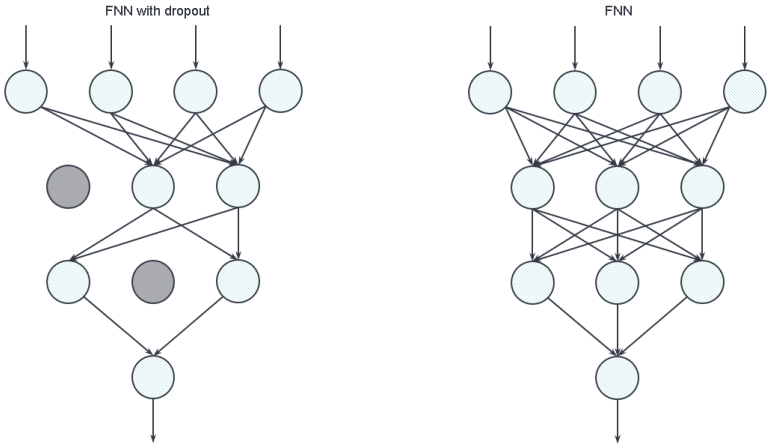

Figure 4.25: Comparison of a neural network with droupout applied during this training step and a general fully conected neural network.

The model architecture must be adapted through the introduction of dropout layers to enable Monte Carlo dropout. However, in general, the computational effort in comparison with bootstrapping can be starkly reduced. Based on this, the Monte Carlo dropout approach applied in this study can be derived: First, each neural network used for point estimation implements dropout layers distributed in front of all hidden layers in the neural network shape Gal & Ghahramani (2016). This is acceptable despite high computational effort and, thus, preserves the ability to apply the gate keeping decision in reasonable time. Second, a random dropout is performed for each of the dropout layers in a random order ten times each. This prevents being stuck in local dropout optima Zhu & Laptev (2017), whereas the number of runs is constrained to reduce the computational effort without starkly compromising on the performance. Lastly, the overall uncertainty can then be derived as follows, where h denotes the number of hidden layers:

$$\sigma^2_{\hat{X}_i} = \frac{1}{10h-1} \sum_{n=1}^{10h} (\hat{X}_{i,n} - \hat{X}_i)^2. \tag{4.28}$$

Empirical methods are required to determine the model uncertainty for neural networks due to the network's complex structure and behavior. The NN prediction error's total variance $Var(e)$, thus, is the sum of the estimated model uncertainty. This model uncertainty is determined with the adjusted Monte Carlo dropout and the variance of the data-inherent noise. Combining this

with a point estimation and the selection of a confidence level, enables obtaining an interval forecast.

4.5 Performance evaluation for prediction and prediction interval benchmarks

To evaluate the applicability of the approach within this section individual subaspects of the proposed intelligent production control for time-constrained complex jobs is evaluated according to the structure presented in Figure 4.26. Therefore, at first the ability of using the foresighted digital twin as a point estimator and its computationally required effort is regarded in Section 4.5.1. Next, Section 4.5.2 regards the state-of-the-art approach and evaluates the point estimation capabilities of the various transitional model approaches proposed. A hyperparameter optimization is performed within. At last, Section 4.5.3 studies the prediction interval approaches that are proposed and identifies the most suitable prediction interval to be used for the final evaluation. Additionally, the model coverage and data coverage are juxtaposed.

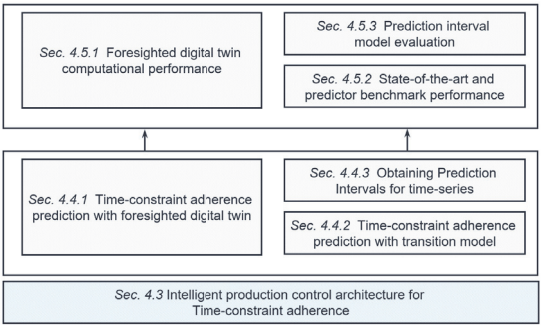

Figure 4.26: Organization of Section 4.5.

4.5.1 Foresighted digital twin computational performance

The computational effort for running a foresighted digital twin as outlined before through an instantiation of a real-world complex job shop of several hundred to more than one thousand equipment is high. Two aspects of computational effort are distinguished as they necessitate certain hardware requirements to be fulfilled. First, the knowledge graph based state of the real-world system at the time of instantiation and its development during the foresight period require storage. Through parallelization this can be multiplied to speed up the simulation process. Experiments show that in the regarded use case 8GB random-access memory (RAM) are more than sufficient to represent such a large system over the course of a rollout.

This allows a parallel rollout on commercially available systems. To preserve industrial know-how exact numbers are not reported but the generalized version in form of the OntologSim (May & Kiefer & Kuhnle & Lanza 2022) for loosely coupled manufacturing systems is publicly and OpenAccess available for anyone to interact and try out. The second computational effort is based on the number of computing operations necessary to run the simulation. The size and implementation play a decisive role as a larger system comes with more events at the same time and the complexity of the transferred production control linearly increases the required computational effort. Based on the available hardware and the overall computational effort the computation time can be obtained. On a more general level the OntolgySim (May & Kiefer & Kuhnle & Lanza 2022) is available to study exact use-case specific computational effort. Again, to avoid the disclosure of details the reported summary can confirm a rollout being possible over several seconds in a commercially available platform. Thus, the gate keeping decision can be based on the foresighted digital twin as long as the number of rollouts is limited or sufficient computing resources are provided. In the evaluation the application to ex post data is less strict with real-time requirements. All in all, the applicability is given with commercially available systems.

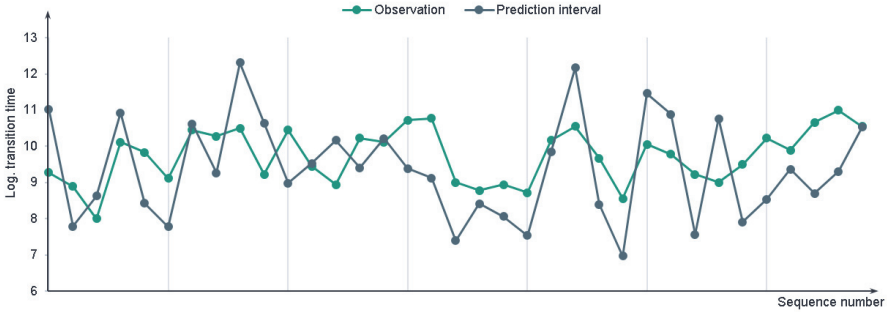

Figure 4.27: Single random sampling rollout point estimator from the foresighted digital twin for an exemplary transition with a sequence of 45 lot transitions.

To measure the prediction ability the foresighted digital twin is used to obtain single transition time predictors from one rollout each. As shown in Figure 4.27 a sufficiently accurate point estimator can be obtained from a single rollout prediction. However, on average the prediction seems to come with a number of clear under- and overestimations. Thus, following the better safe than sorry rationale restricting the lot during a foresight period if only few estimates predict a time-constraint violation is rational. Due to the high computational effort large scale sampling is anyhow hard to implement. Additionally, the foresighted digital twin approach only regards

a small subset of risky or very complex time-constraints, so that this high computational effort is acceptable.

4.5.2 State-of-the-art and predictor benchmark performance

At first, this section introduces the actual real-world semiconductor manufacturing wafer fab that is used to benchmark the predictor performance and represents the common approach of ensuring time-constraint adherence in semiconductor frontend wafer manufacturing as a complex job shop manifestation. Then the state-of-the-art machine learning models used for the single- and multi-variate transition time prediction are introduced.

Industrial time-constraint violation minimization

For the industrial validation a real-world frontend semiconductor fab is used as it represents the most complex job shop with a plethora of time-constraints. Given the technological necessity introduced in Section 2.1.3 time-constraints are heterogeneously distributed on lots and transitions in semiconductor manufacturing. The regarded use case individual equipment in the order of one thousand, each with several process capabilities and machine dedications. As the scheduling and dispatching approaches are not capable of accurately recognizing and incorporating time-constraints on such large system levels scheduling typically ignores the time-constraints. On a dispatching level human operators, each responsible for a work area in the system, can manually perform gate keeping decisions and withhold individual lots if they deem it necessary. Additionally, in some cases manually skipping queues to finally start processing before the end of the time limit can be performed. The latter, however, is hardly ever possible and comes at the expense of other violations and setup increases as well as manual effort. Thus, many time-constraints are violated as the gate keepers are unaware of the overall system situation of this complex job shop.

For instance, in a short validation time slot this error prone handling of time-constraint gate keeping waved a lot through that can barely make it within the time limit. By manually trying to intervene and skip the queues two other time-constrained lots were hindered in adhering to their individual time-constraint. This error prone human decision making is thus a major obstacle. Given shifts and imperfect updates during shift changes the situation is aggravated. Thus, in the regarded complex job shop based on the available data in retrospect any correctly predicted time-constraint violation and gate keeping withholding is a direct improvement over the status quo. Hence, the aim is to build an automated or partially automated gate keeping decision model to reduce time-constraint violations.

Single- and multi-variate predictors

To build such a system the single-variate predictors as introduced in Section 2.3.4.1 based on the pure past transition times are trained and used. To that end the python library *pmdarima* which brings an ARIMA model implementation and automatic p, q, d hyperparameter optimization for the Autoregressive and moving average processes. As ARIMA is capable of expressing at least as good time series in this case as simpler ARMA, AR or MA models, only ARIMA is used. The hyperparameters are internally optimized with a grid search whereas each transition model with its individual time series is optimized. Both p and q are restricted to $\{0, 1, ..., 5\}$ while the order d of the differencing operator is constrained to $\{0, 1, 2\}$ Additionally, a hold-out test dataset is hold and the models can be evaluated. The results of this point estimation with ARIMA can be seen in Figure 4.28

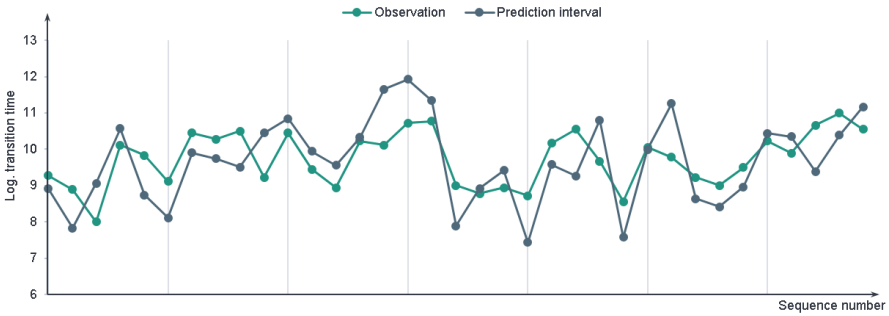

Figure 4.28: Exemplary ARIMA point estimator for a transition.

In a similar vein the machine learning models NN, LSTM and GRU for time-series prediction are trained with the python library *keras*. Multi-variate models are extend with information about current breakdown and queue behavior at the destination equipment following the approach from May & Behnen, et al. (2021). Automation is key as several models, each type once for every regarded transition, has to be built. Thus, the overall procedure is standardized for all models and single-variate to multi-variate models alike. First, an input layer is mapped through one or several hidden layers of potentially individualized layer types to the output layer. In general, a l_2 weight and bias regularization is performed on all hidden layers to improve the Monte-Carlo dropout as proposed by Gal & Ghahramani (2016), that is later used for the uncertainty quantification. For the loss function uniquely the mean squared error (MSE) is used. In order to perform regression tasks MSE is a standard approach with the benefit of the residuals of the time prediction being distributed around zero. This is beneficial for the prediction interval construction. To minimize this MSE as the loss function during the training setting the Adam optimizer is applied to learn biases and weights Bock et al. (2018).

A validation set is used to interrupt training if the validation loss increases for five epochs consecutively. Figure 4.29 visualizes this approach and the result of an exemplary training where the errors converge on both training and validation datasets.

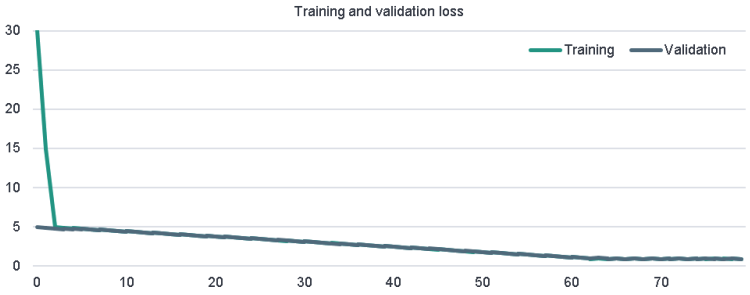

Figure 4.29: Development of the loss for the training and validation datasets over the training epochs.

As for the architectural choices the number of neurons in the first layer is restricted to $\{30, ..., 80\}$, in the second layer to $\{10, ..., 50\}$ and the dropout rate to the interval $[0.1, 0.4]$. As sufficiently large neuronal networks of two or more hidden layers are universal function approximators more hidden layers are excluded for now. These hyperparameters are tuned to identify the best performance on a given metric in the validation dataset. Beyond simple grid or random search that train every possible combination in an inefficient, random, uninform hyperparameter selection in the search space a more guided approach can be used. As systematic, mathematically founded approach bayesian hyperparameter tuning uses a probabilistic surrogate model to predict the score on the evaluation metric in the validation data for input parameter combinations. These predictions and hyperparameters optimized in the surrogate model are the run through an evaluation in the actual model to identify the performance and update the surrogate model. As a python library *hyperopt* offers an implementation of this bayesian hyperparameter tuning used in this case Bergstra et al. (2013). More sophisticated approaches are not regarded as the model size should be constrained to not overly increase the model inherent uncertainty and thus the prediction interval width. A number of different multi-variate data extensions and their encodings are used in the Evaluation and introduced respectively. Nevertheless, the underlying approach remains identical to the herein presented. An example of this is reported in Figure 4.30 and further analyses are available in the appendix.

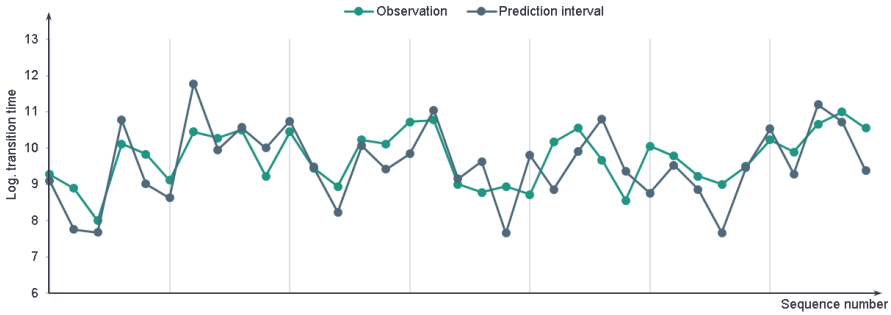

Figure 4.30: Exemplary neural network point estimator for a transition trained on the proposed approach.

4.5.3 Prediction interval model evaluation

Once the individual point estimators are trained and their point estimation optimized the main parameter is the assessment of the interval estimation through its expected coverage $1 - \alpha$. For a given α only the share of α observed realizations of the transition time should exceed the estimated upper bound. There are differences between the intended coverage, i.e. the α selected to optimize the time-constraint violation prediction, and the realized coverage of actually achieved share of transition times that exceeds the estimated upper bound in a post-ex comparison. To do so, a training data to fit the models has to be separated from the validation data used to obtain the realized coverage. As the intended coverage α can be varied a good selection should be achieved.

Note that the overall coverage α is in this case reported and decided for all transitions simultaneously. They could be individually varied and tuned which, however, would result in tremendous manual labor necessary and a high risk of overfit. The results are reported in Figure 4.31 and show a large deviation for the Chebyshev model and increasingly lower deviations between expected optimal coverage and actually realized coverage for the normal and logarithmic model. Thus, while the normal distribution assumption could not be confirmed by the Kolmogorov-Smirnov test as outlined before, the model fit is good enough to significantly reduce the deviations between expected and realized coverage. The logarithmic transformed normal model visually performs best as deviations are smaller. In particular for higher coverages it is important as a coverage smaller than 30% is much less acceptable than a broader coverage to inhibit more time-constraint violations.

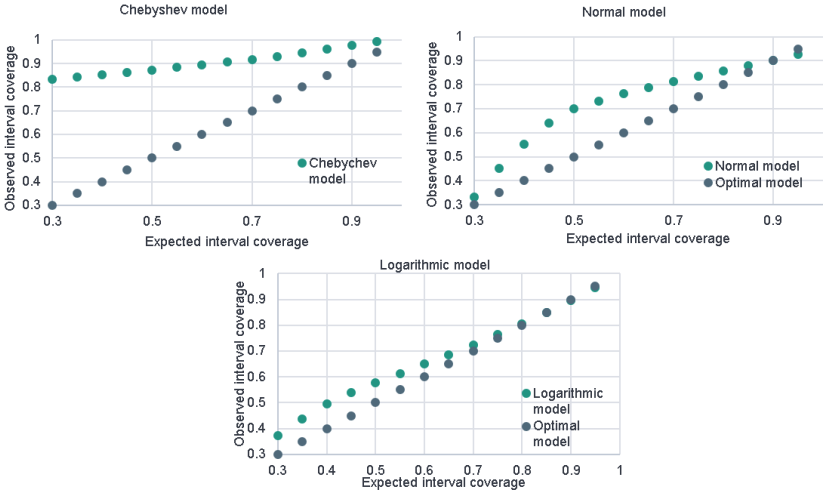

Figure 4.31: Comparing expected and observed prediction interval coverages based on May & Maucher, et al. (2021) and May & Behnen, et al. (2021).

Deriving the optimal coverage level

When challenging the human operators with a coverage level and thus safety level of identified violations higher values are much more anticipated and range between 70% and 100%. As perfect coverage can only be guaranteed by infinitely large prediction intervals this range is reduced up to 99%. The visualization in Figure 4.31 confirms their understanding as in this coverage region the best fit between expected and observed coverage can be seen.

More importantly, the coverage level directly influences the required observations as higher (or lower) coverages can only be sufficiently well assumed if there is sufficient data i.e. observations. This relationship is explained by Handl & Kuhlenkasper (2018) who give the following Equation 4.29:

$$n \leq \frac{2-\alpha}{\alpha} \tag{4.29}$$

Concretely, for instance setting α to 0.05 gives a required minimum of 39 samples as comprehensively illustrated in Table 4.1. As each transition is regarded individually not all possible transitions will fulfill this property and thus not all gate keeping decisions can be modeled. Additionally, simply increasing the observation space hardly helps as this results in longer

observations of time-series and perhaps the underlying reality has much changed since the 39th past observation several months ago. Hence, the desired coverage level and required number of observations must be regarded as a trade-off to be decided against the background of hindering as many time-constraint violations as feasibly possible. Therefore, the sample size and maximum coverage can be limited.

Table 4.1: Comparing coverage levels and required sample sizes for the logarithmic model.

exp. coverage	req. sample size	no. of accept. transitions	perc. of usable transitions
99.00%	199	387	41%
97.50%	79	474	50%
95.00%	39	540	57%
92.50%	26	628	68%
90.00%	19	672	71%
85.00%	13	818	86%
80.00%	9	824	87%
77.50%	8	852	90%
75.00%	7	894	95%
70.00%	6	914	99%

Based on the best performing logarithmic model, as shown in Figure 4.31, for a week long validation time frame 1000 transitions are randomly selected to illustrate the coverage selection process. These 1000 are reduced to transitions that are used at least on two non-consecutive days as otherwise the system behavior is hardly capable of dealing as a prediction for automated decision making. Table 4.1 reports the required sample size for a selection of suitable coverage levels in the interval reported by responsible staff. For each selected expected coverage additionally the share of acceptable transitions is reported. This highlights very well that in the real industrial setting and based on the stringent statistical requirements to close bonds are hardly implementable. Through expert interviews the associated trade-off of coverage and required sample size can be narrowed down to 80% and 75% being acceptable. To achieve a higher number of transitions and still keep a low sample size requirement 80% was identified as the preferred coverage. This will be taken into account for the consecutive evaluation.

Winkler loss prediction interval evaluation

Given the preferred 80% coverage the different prediction intervals approaches are implemented in the above selected validation period. To score the prediction intervals and again compare the possible distributional assumptions the previously derived one sided Winkler Loss is used. Therefore, the period is split into 67% training, the previous validation period,

and the remaining 33% for testing. As no hyperparameters are fit no validation data is required. The predictions and loss function calculation is done on the testing data. The results are reported in Table 4.2. Overall the Chebyshev model performs inferior to the normal and logarithmic model having the highest loss. The distribution-less wide prediction intervals lead to high upper bounds and thus to a high first term loss. The first term loss describes the average loss from the upper bound ($\frac{1}{n}\sum_{i=1}^{n} d_i^u$). In contrast to the logarithmic and normal model the second term ($\frac{1}{n}\sum_{i=1}^{n}[(y_i - d_i^u)\mathbb{1}\{y_i > d_i^u\}]$) is small. As expected the overall loss for the normal and logarithmic model is much smaller with the logarithmic model outperforming the normal model due to lower upper prediction bounds. Thus, the normal models wider prediction interval lead to a larger loss. All in all, for the selected 80% coverage level the logarithmic model is preferred and should be used in the following.

Table 4.2: Custom one-sided Winkler Loss for the regarded period and different distributional assumptions.

Model	First term	Second Term	Score
Chebyshev model	6,844	696	7,540
Normal model	4,434	2,020	6,454
Logarithmic model	3,369	2,70	6,070

Prediction interval for machine learning predictors

Alternatively, if suitable machine learning estimators are used, a hold-out data set can be used to derive the individual prediction interval (May & Behnen, et al. 2021). The disadvantage is that for each model, that is for each regarded transition in this case, a suitably sized hold-out data set has to be taken aside. For simple ARIMA models the prediction of the residual variance can be performed through a separate pass on the hold out data set with the added values of $t_{n-1;1-\alpha}\sqrt{\sigma_{\hat{\epsilon}}^2(1+\frac{1}{n})}$. The newly obtained estimate can then be seen as the constructed corresponding prediction interval (May & Behnen, et al. 2021). For complex multi-variate machine learning models such as LSTMs and NNs the introduced Monte-Carlo dropout is implement to similarly obtain the prediction intervals. A sample is reported in Figure 4.32 wherein the residuals are then used to obtain the prediction interval. As shown the approach is feasible as the residuals are approximately normally distributed and fall around zero.

4.6 Summary of the overall approach and framework

The overall approach for intelligent production control of time-constrained complex job shops is based on a clear model of the regarded production system and proposes a gate keeping production control decision to be implemented in the real system as shown in Figure 4.33.

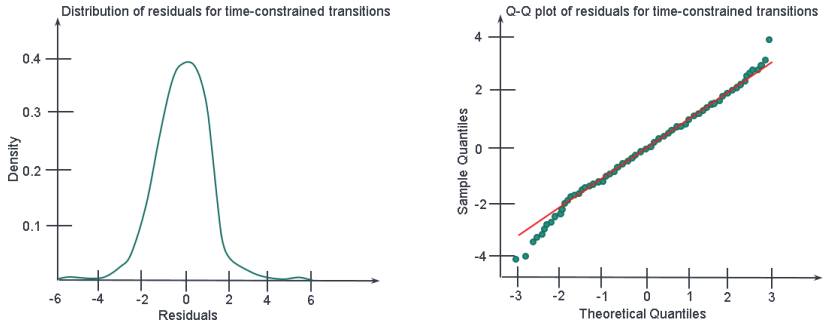

Figure 4.32: Residual analysis and prediction ability of a feed-forward neural network based on May & Behnen, et al. (2021).

From various data sources real-time information about lots and their current place in the production system is combined with general layout and failure and processing information. This fusion can give birth to both a foresighted digital twin in a knowledge graph based simulation and a transition model that exploits statistical relations of these time series so that a good prediction of the estimated transition time can be compared with the time-constraint induced upper transition time limit. A decision model then implements the gate keeping decision and holds back risky lots as evaluated by the uncertainty quantification based prediction interval. Overall, the following process is implemented:

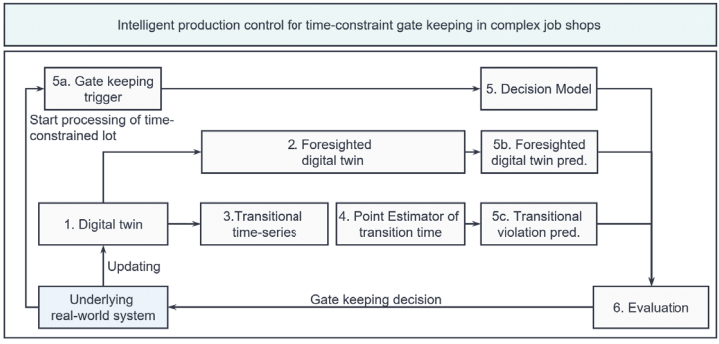

Figure 4.33: Summary of own approach and the implemented process elements.

1. **Digital twin** is the data and data model that is obtained from the up-to-date operation of the complex job shop. It incorporates a data preparation into a knowledge graph based

approach and into transitionally ordered data. This pipeline has to be automated to enable later decision making.

2. **Foresighted digital twin** describes the short term simulation of the digital twin into a period of foresight. It is based on the digital twin as an instantiated and augmented version of the digital shadow.

3. **Transitional time series** are used to describe the typically auto-correlated time series data of transition times based on individual transitions typically between two transitions. This is prepared to create a suitable transition time predictor.

4. **Training of point estimators** describes the fitting of predictors to the individual transition times based on the underlying time series from the earlier step. For each predictor several models have to be fit and optimized automatically. Automating the update mechanism over time ensures that the models are synced with the underlying reality.

5. **Decision model** is the general gate keeping decision model that is based on the prediction interval to quantify the time-constraint violation probability.

    a) **Gate keeping triggers** are used to trigger the decision model by updating the previously described pipeline and querying a gate keeping decision. Typically the possible processing of a time constraint lot is this trigger.

    b) **Foresighted digital twin prediction** is usedif complexer time-constraints or transitions with insufficient data to form an analyzable time series are regarded. Therefore, the digital twin is instantiated and several foresights are regarded based on the defined rollout strategy. This is used to obtain a time-constraint violation probability for the specific lot.

    c) **Transitional prediction** from a trained point estimator as well as the associated prediction interval is performed. As prediction models ARIMA, NN, GRU and LSTM are implemented and used. A comparison will follow based on their operational gate keeping performance. The risk of exceeding and, thus, violating the time-constraint is then derived from student's t-distribution according to $t_{n-1;1-\alpha} = \frac{d^u - \hat{y}}{s\sqrt{1+\frac{1}{n}}}$.

6. **Evaluation** is required to evaluate the time-constraint adherence, implement the gate keeping decision. If the risk identified by the decision model is below the predefined threshold the gate keeping decision does not restrict this lot.

As shown in the performance evaluation all predictors from simple, complex and digital twin based models provide a good point estimator. Similarly, the proposed prediction intervals

fulfill their promise of providing concise upper bounds that can be used to evaluate the time-constraint adherence during the proposed approach. In the following chapter this introduced model is put together and evaluated in an industrial example. Therefore, a semiconductor front end wafer fabs as the most suitable real-world complex job shop serves as the system to be studied.

5 Evaluation and computational results

To properly evaluate the proposed intelligent production control for gate keeping of time-constraints in complex job shops, at first the concretely regarded real-world system use case is described. The description emphasizes the problem setup and available data over the necessary time horizon. To be a suitable validation use case the previously defined requirements are regarded. Then the individual performance of the proposed approaches are reported for the regarded validation use case.

The regarded use case is a semiconductor manufacturing plant which exhibits certain trends and individual peculiarities introduced in earlier chapters. While the overall production planning and control algorithms that operated in a cascaded way are not changed, a gate keeping decision is implemented on the factory level to reduce the time-constraint violations in this exemplary use case. To avoid the trap of building a model and evaluating it on the same available information two longer term periods are regarded. The first for creating and designing the models as well as their initial, already reported, performance evaluation. The second for evaluating the performance as shown in the following. Each separate the time frames into training, validation and test data sets to avoid the same mishap on each model level. These are then used to cope with extreme uncertainty and complexity in the regarded complex job shop. To deal with the frequent changes within such a system a frequent updating and robust approach is needed. Thus, the two approaches based on the foresighted digital twin with foresight and the transitional model are combined. Each is individually and jointly evaluated to reduce time-constraint violations as much as possible without sacrificing general operational performance. Later the overall approach and the general performance summary is presented. The discussion follows in the next chapter.

To that end, the real use case is introduced and discussed in Section 5.1. This includes the setting and transfer as well as preparation required for the applied models. It presents the state-of-the-art performance which serves as the benchmark in this case. Then, Section 5.2 describes the exact performance measures that can be used to rate and compare the performance of individual production control. Next, the foresighted digital twin model is analyzed and its performance reported and discussed in Section 5.3. In a similar vein Section 5.4 analyzes, reports and discusses the performance of the transitional models with its individual underlying single- and multi-variate predictors. Finally, the overall model and performance is summarized and put into perspective in Section 5.5. This includes the selection of the final best performing model combination among the various transitional models that can be combined with the foresighted digital twin model.

5.1 Semiconductor fab as a complex job shop application and benchmark

For validation purposes a second time frame of the regarded complex job shop with a time span of several months is used. The real-world complex job shop is a semiconductor manufacturing factory. Concretely, the frontend wafer fab as the most complex complex job shop available is regarded. Due to the large fluctuations in quantity and actually produced ICs the product mix and state within a semiconductor fab drastically changes daily. As explained the operation on the verge of the physically and organizationally possible makes semiconductor manufacturing the most complex job shop available. This leads to operational excellence being the single most important target for semiconductor manufacturers. Thus, having an efficient production planning and control system that highly utilizes capital intensive equipment and ensures technological required constraints such as time-constraints are adhered to is paramount. Therefore, regarding such a semiconductor fab for validation purposes is ideal. To avoid revealing sensitive data certain values over time and sizes are not fully reported, the time-constraint adherences however can be revealed.

This regarded wafer fab has a large number of equipment in the magnitude order of one thousand. Thus, the number of potential transitions is in the magnitude order of one million. However, the actual number of observed transitions is less than 5% of these. Of these only a small subset has time-constrained lots that travel on the respective transition. Additionally, not every lot that travels on such a potentially time-constrained transition has an upper time-limit due a time-constraint. Overall, in a subset of the regarded validation period there were approximately 10.000 time-constraint transitions on the equipment ranging in the order of thousand. Of these time-constrained transitions significantly less than 5% were violated. While this is a far too low number to provide any end to end learning of violation or adherences with such an imbalanced dataset, the actual number of violations being far greater than 100 leaves great room for improvement. As any lot with its wafers is associated with a value in the magnitude order of a new car (Mönch & Fowler & Mason 2013) there are significant economical savings possible. Likewise, ecological savings are possible as the energy and water usage for semiconductor manufacturing is among the highest.

All in all, over the regarded period each time-constrained lot that is expected to exceed the time-constraint prescribed time limit has to be withhold by the gate keeping decision. For one exemplary transition the actually observed transition times can be seen in Figure 5.1. Clearly, the time series is highly complex. To evaluate the performance of the proposed model in being able to correctly predict time-constraint violations this historic data is regarded and observed time-constraints are ex post evaluated with the proposed approach. If the approach is capable of identifying time-constraint violations observed in the real-world data

it can offer superior performance. Therefore, the current human operator based handling of time-constraints which is indirectly observed in the data serves as the benchmark. As introduced in the state-of-the-art other approaches are not capable of handling such large scale semiconductor manufacturing systems as the complexity explosion leads to excessive computational effort.

Figure 5.1: Exemplary observed transition time and split into training, validation and test time frames.

Additionally, when regarding both the foresighted digital twin model and the transitional modeling a sufficiently large past behavior has to be observed to recreate the actual system states. This includes data about maintenance and breakdowns in the magnitude order of several ten thousand events during the regarded period. Thus, the classic split of the data into training data, to create the data basis and train the parameters of prediction models, validation data, to select the most suitable models with hyperparameters based on previously unseen data, and test data to evaluate the final performance is applied. The training data is set to 70% while validation and test data each make up 15% as illustrated in Figure 5.1.

5.2 Performance evaluation of time-constraint adherence

To evaluate this proposed model on ex post data several performance metrics can be used. First, Section 5.2.1 introduces performance metrics that can be used to evaluate and compare the prediction intervals from different models. This can help comparing the differently obtained and different types of prediction intervals. Additionally, individual point estimators can be evaluated with traditional errors such as the mean squared error (MSE). However, this point estimator evaluation itself is meaningless, as the overall performance in correctly providing a decision for the gate keeping decision for time-constrained lots is what really counts. Irrespective of the model used, irrespective of the prediction interval and decision logic, it all

boils down to how many time-constraint violations can be prevented. In the regarded ex post evaluation of the data set this is equal to the number of correctly predicted time-constraint violations in the dataset. Section 5.2.2 provides the performance metrics for this evaluation.

5.2.1 Performance metrics for the evaluation of Prediction Interval Quality

For the evaluation of the prediction interval quality both the prediction interval coverage probability (PICP) and the mean prediction interval width (MPIW) are required (Khosravi et al. 2011). The PICP should exceed the confidence level $1 - \alpha$ as it is constructed through the division of the number of predictions not exceeding the upper limit d^u through the test set size n_t. It's importance stems from the fact that the PICP measures the ability of the prediction interval to properly classify the upper limit exceeding realizations according to the following equation.

$$PICP = \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{y}_i, \text{ with } \hat{y}_i = \begin{cases} 1, & \hat{y}_i \leq d_i^u \\ 0, & \hat{y}_i > d_i^u \end{cases} \tag{5.1}$$

Second, the width of the interval has to be controlled to avoid a perfect coverage with arbitrarily wide prediction intervals. This trade-off is found in the MPIW which can be calculated as the sum of upper bounds divided by the number of predictions as follows.

$$MPIW = \frac{1}{n_t} \sum_{i=1}^{n_t} d_i^u \tag{5.2}$$

Comparing Multiple Models

To compare the prediction intervals obtained from multiple models the absolute Error (AE) can be used. The AE is the sum of the predictions' absolute residuals calculated with the observation y_i and model prediction $\hat{y}_i$. One alternative to the proposed model is a simple predictor which in any case predicts the mean of the observations $\overline{y}$. To put the model's performance into perspective the relative absolute error (RAE) can be used, which compares the model to the simple predictor's performance as formalized in the following. The benefit of the model, compared to the simple predictor, can also be used for comparison as in Equation 5.3.

$$RAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \overline{y}|} \tag{5.3}$$

5.2.2 Performance metrics for the binary classification evaluation

The evaluation of time-constraints is based on a binary classification, i.e. whether or not a lot has exceeded the maximum time allowed. Recall, precision and accuracy, as defined in Table 5.1, are used to evaluate the model performance. These are based on two characteristics for each time-constrained lot, first the binary value of violation (positive) or adherence (negative) to a time constraint and second the binary value of the model classification (true) or misclassification (false). The resulting four classes are the number of True Positive (TP), where positives stands for violation and true for being correctly classified by the model, True Negative (TN), correctly classified adherences, False Negative (FN) and False Positive (FP).

In short, recall is defined as the division of correctly identified violations by all predicted violations and interpreted as the proportion of correctly identified true positives. Precision in a similar vein can be interpreted as the faith one can have into the validity of a violation prediction. As a measure of the overall prediction ability, independently of violation and adherence, the standard criterion of accuracy is used.

Table 5.1: Recall, precision and accuracy calculation.

Recall	**Precision**	**Accuracy**
$\frac{TP}{TP+FN}$ 5.4	$\frac{TP}{TP+FP}$ 5.5	$\frac{TP+TN}{TP+TN+FP+FN}$ 5.6

5.3 Evaluation of Foresighted Digital Twin-based production control approach

In the following the foresighted digital twin based gate keeping production control approach to achieve time-constraint adherence is evaluated. The evaluation takes place on the 15% test data so that a comparison to the later evaluated machine learning based prediction models is fair. The preceding behavior as captured in the data captured before is used to create the up-to-the-minute representation of the real system used to instantiate the digital twin as outlined before. However, as to the large computational effort of the digital twin model and

as to follow the previously introduced decision model the digital twin and its foresight period should only be used for unclear and complex situations. Thus, the selection of time-constraints is reduced to frequent violators and complex time-constraints. Otherwise the model complexity would not enable acceptable runtimes. At first, the simulation and its prediction capabilities are evaluated in Section 5.3.1. Next, Section 5.3.2 evaluates the performance in the binary time-constraint adherence through gate keeping task.

5.3.1 Evaluation of the simulation model prediction

As shown in Section 4.5.1 a single random sampling rollout point estimator can provide a prediction for the next transition time. While errors are larger than with perfectly fit prediction models the foresighted digital twin has other values. To evaluate the simulation model and its capability of generally predicting the transitions times sufficiently well a larger evaluation has to take place.

Therefore, for a time-constrained transition with violations the random sampling rollout point estimation was repeated for 20 times to obtain 20 different point estimators under different stochastic drawn samples. Note, that this process is very resource intensive so that an overall evaluation of the entire dataset over months for all transitions is impossible to obtain. Nevertheless, sufficient performance over changing environments over time for the complex time-constraints is a fair evaluation. The foresighted digital twin model is based on simulating the overall wafer fab with processing, material flow and breakdowns and maintenance. Thus, for each rollout within the knowledge graph based digital twin the entire fab is simulated for a certain foresight period. Due to the computational complexity an application to any time-constraint in the real world data set is hardly possible with limited resources. However, the major advantage is that the foresighted digital twin does not need a minimum number of observations as time-constraints and can be applied for complex time-constraints. Therefore, it is not necessary to evaluate the simulation as a prediction model on all transitions over the regarded horizon, but sufficient to evaluate its performance on a subset.

The results of this evaluation are presented in Figure 5.2, where the time limits from the time-constraints and the observed transition times are reported. The point estimator of the predicted transition time is obtained as the mean of the 20 transition times obtained from the rollouts. In a similar vein, this discrete residual function is used to present a cutoff of 80% mimicking the prediction interval. 80% are selected as the maximum upper bound as illustrated in Chapter 4.

Overall the prediction is visually seen quite accurate and the observed transition times for a large majority fall into this 80% confidence interval. Remarkable is that the fine-tuned simulation through the regarded number of rollouts does not seem to have a bias towards

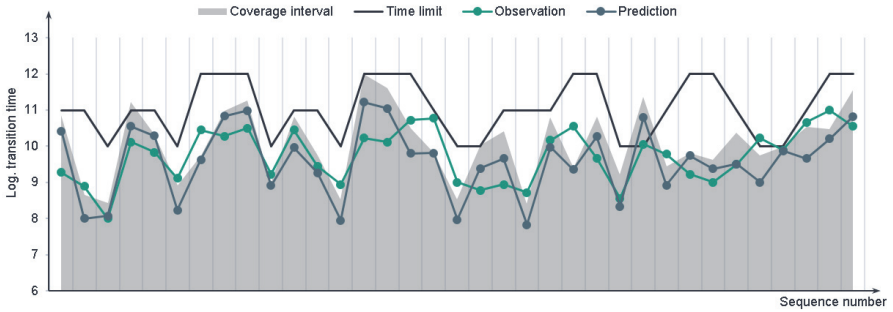

Figure 5.2: Exemplary observed transition time, predicted transition time, time limits and the prediction interval at the designated coverage, where actual violations can be seen whenever the observed transition time exceeds the time limit.

over- or underestimating the transition times. This behavior is often seen in regular simulation projects. However, they do not rely on instantiated digital twins but rather averages and normally regard different performance characteristics than the transition time. Nevertheless, in quite a few number of observations the observations deviate from the selected interval. Thus, this confirms that for the gate keeping decision for time-constrained lots a more conservative approach must be taken. Therefore, 5% of observed time limit exceeds are taken for the later binary evaluation of the capability to withhold time-constraint violators. In case 5% or more of the observed realizations in the rollout violate the time-constraint the lot is restricted.

5.3.2 Evaluation of binary classification

As explained the foresighted digital twin and its prediction capabilities heavily depend on the rollout and decision logic that takes into account this rollout strategy. Due to the complexity a rollout strategy evaluation is not possible. Therefore, based on the results shown above, the *better safe than sorry* strategy is used in which the rollout is a seeded random sampling and the decision logic is the simple heuristic of only allowing lots to be picked for processing that have not shown a 5% or higher chance of violating its time-constraint within the observed rollouts. Based on the reduced evaluation data set used for this foresighted digital twin to preserve computational efficiency and focus on the complexer time-constraints the following confusion matrix can be reported in Table 5.2.

Of the regarded time constraints, the vast majority has in the real world system adhered to the time-constraint. This regarded subset, due to the selection rules of complex and critical time-constraints, is not representative. 21 time-constraints were violated which is below 2%. Out of these more than 65%, known as the recall, were correctly identified and could have

Table 5.2: Confusion matrix for the binary classification of time-constraint violation and adherence of the foresighted digital twin model in the reduced test data set.

	Actual violation	**Actual adherence**
Predicted violation	14	458
Predicted adherence	7	2011

been saved from scrapping with the proposed approach. In general recall should be aimed at reaching higher values of 95% or more. But that would inevitably increase the number false positive predicted violations of actual adherence and hence lead to a much lower precision than the achieved 3%. The achieved 3% is from a general perspective low. However, one has to keep in mind that the costs of a false negative prediction are by several orders of magnitude larger than that of false positive gate keeping decisions. The latter is acceptable as long as equipment utilization can be kept high and the lot that is hold back can start processing in one of the subsequent decisions. Thus, the result is a great improvement over the state-of-the-art production control that served as the comparison by providing the underlying data set and reached 0% precision. Overall, the model achieves an accuracy of approximately 81%, as shown in Table 5.3.

Table 5.3: Recall, precision and accuracy for the foresighted digital twin based approach.

	Foresighted digital twin approach
Recall	66.67%
Precision	2.93%
Accuracy	81.33%

5.4 Evaluation of transitional model based production control approach

In a similar vein to the evaluation of the foresighted digital twin for gate keeping production control for time-constrained complex job shop the transitional modeling approach is evaluated. The underlying dataset is identical and the 15% test dataset is used for pure evaluation. However, as the transitional modeling approach cannot deal with complex time-constraint these are excluded. Additionally, the dataset covers all regarded transitions and hence a larger number of transitions as the foresighted digital twin approach as the computational effort is far lower than for the knowledge graph based simulation. At first the selected predictors and the architecture and hyperparameter optimization, if applicable, are briefly stated. Then, Section 5.4.1 presents the evaluation of the point estimators and prediction intervals. The final binary classification of the gate keeping production control performance to ensure time-constraint adherence is presented in Section 5.4.2.

Predictor models used

Overall, the following machine learning prediction models are trained and used to obtain the point estimators and prediction intervals. First, an ARIMA model is fit to the time series followed by a feed-forward neural network. Then, GRU and LSTM models are fit. They are briefly mentioned in the following.

First, as a simple time series model an ARIMA model is selected. As shown by May & Maucher, et al. 2021 the increasing capabilities of more sophisticated simple time series model like ARIMA and ARMA greatly outperforms the simple AR and MA models. Thus, it is sufficient to regard the most sophisticated approach. Due to the low number of parameters to be fit the model is still easy to handle and fit.

To increase the ability of incorporating non linear relationships a neural network is fit. This NN is based on the proposed multi-variate approach from May & Behnen, et al. (2021). Therefore, it incorporates current breakdown and queue behavior at the target machine. Again, a good compromise between many parameters and possibly better point estimators but much wider prediction intervals has to be made.

Similarly, the LSTM approach is based on the proposal from May & Behnen, et al. (2021). As the LSTM is made to be fit to spatio-temporal data such as the regarded time series of transition times it has performed well in that study. The multi-variate approach is identical to the NN to achieve both good point estimators and prediction intervals.

Lastly, GRU are novel time series machine learning models that provide a good fit for various time series (Yamak et al. 2019). In other applications their results have been on par or better than state-of-the-art models such as LSTM.

All models are trained and the prediction intervals obtained with previously outlined methods. They are automatically fit as introduced later.

Architecture of the ML Models

The model architecture and approach is identical to the architecture and hyperparameter search introduced in Section 4.4.2. As explained for every transition, the model input is derived and a model is trained. A good model fit is irreplaceable in using it to obtain good point estimators and prediction intervals. However, the quality cannot be manually improved for the large number of transitions that is to be regarded.

Each model's architecture and hyperparameters are optimized according to the following procedure with a fixed training, validation and test data set split of 0.7, 0.15 and 0.15.

- Architecture: A variety of experiments if performed on the number of layers, bias and kernel regularization as well as activation function for the feedforward neural networks with minimum two dense layers.
- Sequence length: Through a grid search the best performing length of the input sequence is selected.
- Hyperparameters: The optimal hyperparameters are found with a grid search implemented as the python library *hyperopt*.

5.4.1 Evaluation of the influence of multi-variate prediction intervals

After selecting the coverage and method to calculate the prediction interval the individual models can be built. ARIMA is an uni-variate time series modeling approach that can solely predict the transition time based on observed past transition times. All other, namely NN, LSTM and GRU are multi-variate models that can incorporate more parameters. May & Behnen, et al. (2021) have studied the influence of further parameters on the transition time prediction capability and selected the models that best improve the decision. Within this thesis the multi-variate selection is based on that study so that NN, LSTM and GRU contain current breakdown and queue behavior at the destination equipment. As the prediction has to be made in advance no future queue information is available and hence cannot be integrated. To address such an issue May & Albers, et al. (2021) successfully predicted future queue lengths based on current complex job shop state information. However, the approach still yields an error larger than 5% which diminishes their explanatory power concerning future transition times. Also, more parameters increase the neural network sizes which lead to wider prediction interval which is undesired.

Thus, the prediction intervals of these multi-variate models can be compared with the introduced performance measures prediction interval coverage probability (PICP), mean prediction interval width (MPIW) and relative absolute error (RAE). Table 5.4 reports these values. Clearly, the PCIP close to 1 and equal or higher than the selected coverage interval of 80% favors LSTM models. Note that this is averaged over the whole dataset. Likewise, MPIW and RAE show better prediction interval performance for the LSTM model. Due to the LSTM approach the number of parameters in its structure with a similar architecture is lower than that of a NN for instance, which could explain the lower epistemic uncertainty and thus better prediction intervals.

Beyond the purely statistical comparison each model with the maximum coverage of 80% can be illustrated for exemplary transitions to visually confirm the good point estimators and prediction intervals. An example transition at a random point in time of the validation data

Table 5.4: Prediction interval coverage probability and mean prediction interval width of the NN, LSTM and GRU models.

	NN model	GRU model	LSTM model
PICP	0.8941	0.9256	0.9390
MPIW	9.677	8.977	7.671
RAE	0.9996	0.8720	0.6957

is selected and reported with time limits from the time-constraints for the ARIMA model in Figure 5.3, for the NN model in Figure 5.4, for the LSTM model in Figure 5.5 and for the GRU model in Figure 5.6. Clearly, all models in general can be used as good transition time estimators that visually correspond both for point estimators and the associated prediction intervals to the later observed real data. The time-constraint adherence can thus be predicted very well. Notice that in the NN model a few false alarms would have been triggered with the 80% coverage interval. Thus, in the following for the binary classification the preferred 80% coverage interval is used.

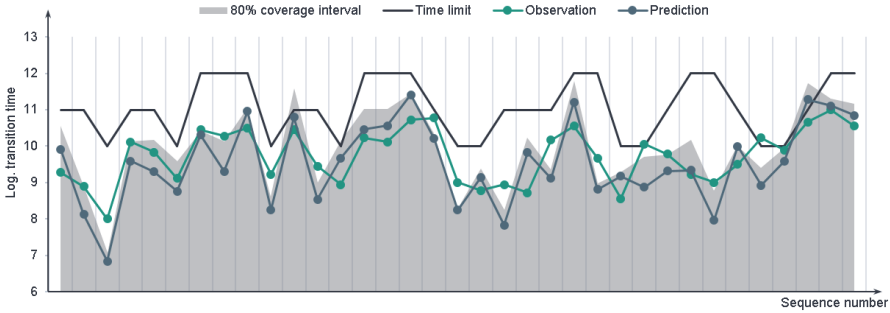

Figure 5.3: Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with an ARIMA model.

All in all, the proposed uni- and multi-variate prediction models perform sufficiently well by providing acceptable point estimators and small prediction intervals. Overall, the ARIMA and LSTM model outperform the NN and GRU model with respect to the regarded prediction intervals with coverage or confidence of 80%. Nevertheless, all models are later evaluated against the binary classification as the pure statistical and visual evaluation is insufficient.

5.4.2 Evaluation of binary classification

Implementing the transitional models with the preferred 80% coverage level enables a direct comparison of their performance. To do that the models were implemented and used the model

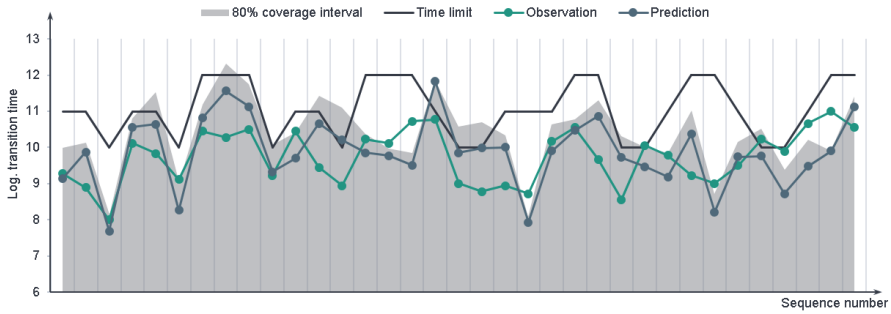

Figure 5.4: Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a NN model.

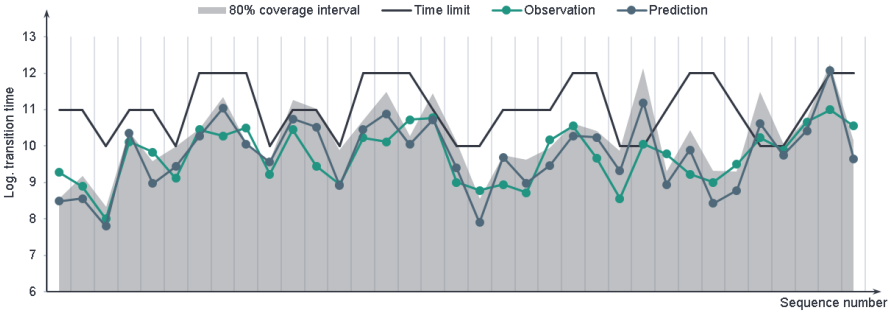

Figure 5.5: Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a LSTM model.

inherent epistemic uncertainty estimation techniques that showcased a better performance for the prediction interval for machine learning models. For the ARIMA implementation the preferred logarithmic transformation of a normal model is used to derive the prediction interval. Then the identified performance metrics are reported in Table 5.5.

Recall describes the ability to detect the violated time-constraints. In this performance metric the simple ARIMA based model performs best and can identify and correctly withhold all time constraint violations. The LSTM model is able to withhold 90% correctly which is its closest match. Both GRU and NN model perform inferior.

The precision shows the performance of correctly withholding lots without high numbers of unnecessarily restricted lots. This performance measure is secondarily important as the utilization of capital intensive equipment has to be high. Here, LSTM performs best but

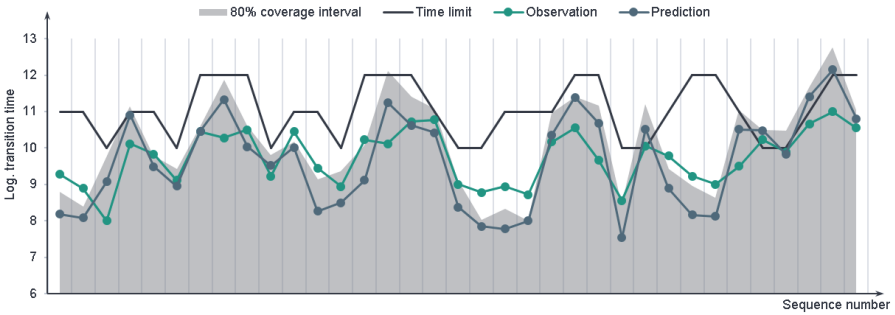

Figure 5.6: Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a GRU model.

Table 5.5: Recall, precision and accuracy comparison for the selected prediction models in the transitional modeling approach with the preferred 80% coverage interval.

	ARIMA model	NN model	LSTM model	GRU model
Recall	100.0%	38.46%	90.9%	82.9%
Precision	11.73%	2.82%	16.67%	3.78%
Accuracy	97.31%	95.21%	98.72%	96.5%

still restricts about six times the number of necessarily restricted lots. The ARIMA model performs only slightly less at about 12%. Both GRU and NN model cannot keep up with this performance.

From an overall perspective using the accuracy yields good performances of all models in the magnitude order of 95% to 99% which is a very good performance from a machine learning perspective. However, as the dataset is highly imbalanced and most time-constraints in the observed time period are processed in time a simple *always predict adherence* would perform comparably or even better. Thus, it is important to use domain knowledge to build and evaluate the models, in this cased centrally on Recall and Precision.

All in all, both the simple ARIMA model due to its simplicity and capability of avoiding overfit and the complex LSTM model that can handle long- and short-term data perform best and can withhold almost all violated time-constraint. Monetarily that would translate to saving in the magnitude order of million Euros for the regarded time period.

Nevertheless, there are still too many requirements and several false negative predictions that impede a direct implementation with industrial companies. For an industrial implementation the reliability of the digital twin interconnection to the underlying real-world data has to be

ensured. Likewise, the change management driven decision about the exact implementation form and logic has to be taken. This includes in particular in which way the gate keeping decision is implemented to the responsible worker. In general weak indications, warnings for potential high risk selections up to prohibitive measures are possible. Lastly, the internal financial evaluation has to take place to ultimately weigh up the savings from avoiding time-constraint violations with the hidden costs of false positive, i.e. falsely restricting lots. Based on the preliminary evaluation these costs are currently negligibly small. However, for a holistic evaluation implementation costs and costs for keeping this intelligent production control system up to date as well as all further costs and savings have to be regarded in an integrated manner. Furthermore, comparison with alternative projects cannot be overlooked. To address this gap further research is necessary, for instance improving upon hyperparameters for an individual transition as visualized in the appendix.

5.5 Summary of evaluation and computational results

In a nutshell, the evaluated foresighted digital twin and transitional model based intelligent production control for time-constraint adherence in complex job shops lives up to the promise of minimizing the number of time-constraint violations. In the regarded validation real-world semiconductor wafer fab more than two third of all time-constraint violations could have been prevented by the proposed model. Considering the value of such scraped lots in the magnitude order of several ten thousand Euros, this would relate to financial savings in the magnitude order of millions of Euros. Moreover, the large stress on operators could be reduced and valuable expert attention focused on different aspects. Additionally, that reduced scrap could massively reduce the carbon and water footprint of produced chips.

Concretely, simpler time-constraints can sufficiently well be classified with the fast transitional model. Here surprisingly both the simple ARIMA model due to its low complexity and epistemic uncertainty as well as the LSTM model due to its good point estimators and acceptably small prediction interval perform best. Nevertheless, the NN and GRU model still perform acceptably well as they could still save a considerable share of the scrapped time-constrained lots in the validation wafer fab and time frame. Complex time-constraints can be actively controlled through the same gate keeping production control decision by using the foresighted digital twin based approach. The computationally intensive foresighted digital twin model can still be used to control complexer time-constraints and selected periods. It performs acceptably good and is able to withhold a large share of time-constrained lots that would otherwise have to be scrapped. Even for these complex time-constraints more than half of the time-constraint violations would have been avoidable with the proposed approach which cannot be controlled with any traditional approaches.

Therefore, the overall model is capable of avoiding a large share of time-constraint violations that are currently unavoidable and cannot be reasonably controlled. Concretely, between 67% and up to 99% of the time constraints that were violated in the real world setting could have been prevented. Consequently, the evaluation and real-world data evaluation can be regarded as a success. Likewise, computational complexity is insofar limited as the applicability to real world cases could be shown.

Thus, all in all the evaluation of the proposed intelligent production control for gate keeping of time-constrained lots in complex job shops yields competitive results superior to implemented state-of-the-art approaches. Therefore, an implementation into real-world complex job shops should be regarded in subsequent work.

6 Discussion and Outlook

An approach for intelligent gate keeping production control of time-constrained lots in complex job shops is presented within this work. Based on a real-world semiconductor frontend wafer fab as the most complex job shop the approach is validated. Section 6.1 juxtaposes the approach and its computational results with the research questions, research deficit and requirements derived in Section 1.3 and Section 3.2 as well as Chapter 2. A future research direction is presented in Section 6.2.

6.1 Discussion

Deriving an intelligent production control for time-constrained complex job shops based on real-time data is the underlying research goal of this work. On the bedrock of this overarching goal five research hypotheses are set forth. Corresponding requirements were derived and the state-of-the-art was evaluated to extend this to the overall research deficit. Based on the research questions the approach and results are critically discussed in the following.

1. How to use static and dynamic knowledge graph based production system replicas to support production planning and control with time-constraints?

The implemented foresighted digital twin is based on a knowledge graph based production system replica. Both static information about machines, capabilities and layouts as well as dynamic information about the current system state in form of the position of individual lots in the system and the maintenance and breakdown behavior are intertwined. Through the current state information the digital twin is instantiated to either accurately train and derive the following production planning and control decision through foresight or to evaluate the behavior control behavior in a more real-time like manner. The former is implemented and used within the proposed approach to control the gate keeping decision to improve time-constraint adherence. As illustrated in the results the approach is applicable in complex cases that cannot be solved otherwise. The results show that such knowledge graph based system replica helps improving the control system on the desired target. Time-constraint adherence can be more accurately controlled with the help of this knowledge graph based simulation than with the state-of-the-art methods.

However, the computational effort for such an highly detailed knowledge graph based simulation is very high. Therefore, results are only obtained after a certain time period. Additionally, despite the graph based approach it remains a discrete event simulation that cannot be analytically described or derived to find optima. Moreover, the required data and information level is very high and augmentations are necessary to successfully instantiate this foresighted

digital twin. While automation can alleviate some of this pain a large effort is still required. Nevertheless, due to the implementable reasoning and automation the effort can be kept lower than for alternative production system simulation models.

2. How to use real-world real time data to avoid time-constraint violations with a data-based approach for production control for complex job shops?

The presented approach to minimize time-constraint violations and hence to avoid as many time-constraint violations as possible can be based on available real-time data used in time series models. On each individual transition which is time-constrained an individual model for the time series of time spent between ending the process on the source equipment and beginning the subsequent process on the destination equipment is built. A number of time series modeling techniques is then used to predict a point estimation of the required transition time. This transition time prediction can then be evaluated against the known time-constraint time limit. As shown the approach in general is feasible to obtain good transition time predictions and facilitate a decision model. To accurately reflect the uncertainty in prediction a prediction interval is used during this decision model. The results show that the majority of state-of-the-art uncontrolled time-constraint violations can be avoided with this real-time data based approach.

However, this data intensive approach requires a large number of observations and hence for a complex job shop large data bases have to be held up to date. In turn, each time-constrained transition time series needs one model to be fit, stored and evaluated as well as retrained frequently to incorporate novel data and concept drifts. Thus, a non negligible effort for keeping and using these models has to be made. Nevertheless, this effort for the regarded time span of the real-world semiconductor manufacturing wafer fab was tolerable by commercially available hardware. Therefore, implementations are still possible, even for large complex systems yet they are not to be underestimated.

3. How to enrich and extend machine learning algorithms to accurately capture the aleatoric and epistemic uncertainty in large-scale complex job shops when predicting time-constraint adherence?

For the above presented time series predictors several complex state-of-the-art machine learning algorithms are used to obtain point estimators. To include the uncertainty stemming from the complex interconnected material flow, frequent control changes and maintenance which heavily influences the stochastic distributed transition times a feasible approach is presented. First, the overall aleatoric uncertainty and distribution of these transition times can be used to obtain a prediction interval based on assumptions on the underlying distributions

which are then fit. Through the novel proposed one-sided Winkler loss they can be evaluated. Second, alternatively the uncertainty can be derived from the machine learning model inherent, epistemic uncertainty that can be evaluated through Monte Carlo dropout on a hold-out dataset. The latter approach requires higher computational effort and leaves less observations for the initial fit as the hold-out dataset needs to be reserved. Both approaches can be effectively used to derive the prediction interval of the transition time around the point estimator. Combined these can be compared to the time-constraint inflicted upper time limit to derive the time constraint adherence probability. The results show that this probability can be accurately predicted ex ante so that the gate keeping decision can greatly reduce the number of time-constraint violations.

However, process has stringent requirements on the minimum number of observations to ensure sufficiently small prediction intervals. Therefore, it is not applicable in few rarely observed cases. Additionally, complex time-constraints cannot be effectively evaluated based on these transition time predictions alone. Nevertheless, the vast majority of time-constraints can be controlled and most violations can be prevented with this fast applicable real-time method. Still, the computational effort for obtaining these distributions and uncertainties should not be underestimated. Frequent changes in a complex job shop require constant updating.

4. How to use long-term and real-time knowledge acquired within a factory to holistically reduce time-constraint violations with intelligent production control?

To sufficiently address this first research question the focus is put on complex job shops which signify the most complex discrete production systems and include time-constraints as the most adverse condition. The model of a complex job shop is general enough to cover a wide range of flexible production system even with less complex circumstances. Thus, the complex job shop model is well suited cover nowadays complex and less complex systems as well as their increasing complexity in wake of the mega-trends faced by manufacturing. Explicitly, production planning and control on a system level and through state-of-the-art implementations is considered. As the complexities culminate an approach to control the gate keeping of lots as part of production control is derived. This approach makes use of long-term data and knowledge of previous behavior and layout information distilled into the transition models and foresighted digital twin. Additionally, real-time data is implemented into these decision models to facilitate real-time decision making for improving the time-constraint adherence. However, this reliance on historic behavior and real-time data permits the immediate application in novel circumstances as ample time to learn and adapt is required. Using the foresighted digital twin approach can shorten this adaption and learning time for

novel systems and bridge concept drifts. Nevertheless, there are minor limitations in the applicability to any novel production system using the provided framework.

The underlying approach to facilitate these gate keeping decisions is based on a twofold time-constraint adherence evaluation. First, a current state based foresighted simulation model is used to obtain a transition time prediction through simulating through the foresight period. Second, a transitional modeling approach uses single- and multi-variate time-series models to predict future transition times. These predictions can be evaluated against the prescribed time limits by using prediction intervals and optimized decision rules based on the obtained time-constraint adherence probability selected according to the coverage. This framework is generalized and can easily be transferred as it is based on dynamic decision making. There is a high compatibility to the manual state-of-the-art control and rule-based approaches. This dynamic and interacting production control is adaptive i.e. reacting to the system and its likely future behavior.

Knowledge is acquired through the application of learning algorithms on historic and current data. This distilled knowledge is available through the machine learning models which implement intelligent behavior. Therefore, the intelligent real-time production control for complex job shops is realized. The results show that the majority of currently not identified time-constraint violations can be effectively controlled and avoided with the proposed approach. Overall the models greatly outperform the current production control methods used to control time-constraint adherence as shown in the real-world semiconductor frontend wafer fab validation.

However, as shown in the overall results there is trade-off between the coverage of identified time-constraint violations and the number of unnecessarily withhold lots. Thus, careful fine-tuning and implementation with a large degree of domain knowledge is required.

5. How does the learning-based intelligent production control for complex job shop perform in ensuring time-constraint adherence in a real-world setting?

The proposed approach is validated in a real-world semiconductor manufacturing frontend wafer fab which serves as the most complex real-world complex job shop. Complexities and time-constraints are abundant, so that the number of observed time-constraint violations is significant. Yet, the state-of-the-art production control is already capable of preventing the vast majority of time-constraint violations. Thus, the proposed approach which makes use of transition time predictions with machine learning and a digital twin based on all observed transition times irrespective of time-constraint violations has a decisive advantage over any end-to-end time-constraint control approaches. Over two separately regarded datasets containing the real behavior of each several months the approach could prevent the majority

of time-constraint violations of the state-of-the-art methods. Thus, it vastly outperforms its benchmarks and could save scrapping of lots worth of in total up to several million Euro.

However, the performance decreases with the time-constraint complexity as the uncertainty increases vastly. Moreover, a large number of false positive predictions shows that also non-violating lots are withhold unnecessarily during the gate keeping decision making. Up to a certain degree this is acceptable as the costs of withholding are negligibly small compared to a time-constraint violation. Nevertheless, decreasing the number of false positive violation predictions without compromising on the true positive violation predictions could improve the models further and spur faster integration. Additionally, the approach is currently validated with the real-world wafer fab data but not implemented and used by operators as a real-time control system. For the validation regarding the same system during another time-frame is sufficiently showing the applicability, but for reaping the benefits of the approach an implementation and further fine-tuning would be beneficial.

6.2 Outlook and further considerations

A contribution towards the research deficit of effectively controlling time-constraint violations through production control in complex job shops is made through the presented approach. This approach takes a great step towards better production control for complex job shops through sophisticated, artificial intelligence based models. There is still a long way to go and several future research directions were identified through the course of this work.

First, the presented approach is implemented and validated on several months of a semiconductor manufacturing plant. This frontend wafer fab is the most complex available complex job shop. Nevertheless, due to the vast number of possible transitions and events some are still only rarely present in the dataset leading to possibly improvable performance. Therefore, more data from more systems and from longer time frames should be regarded. Additionally, the implementation of long-term fit simulations which learn the simulation fit to reality for a full wafer fab and could provide artificial training data. Similarly, a coupling of these approaches with the knowledge graph based digital twin could speed up its computation improving results and applicability. Alternatively, these rare events and missing data based transitions could be regarded with techniques that reduce the required data such as cross validation.

Second, the currently used machine learning models apply a loss function that optimizes for the point estimator but does not regard the prediction interval directly. Thus, including the prediction interval into the loss function could enable the machine learning model to perform better on this multi-job task of good point estimators and narrow prediction intervals. For deep neural networks Jiang & Deng (2023) propose an approach to include confidence intervals which could be evaluated for extension towards prediction intervals. Additionally,

the performance of the multi-variate predictors could be improved through extensions with more comprehensive features. For instance, a fusion with the foresighted digital twin to more accurately reflect the system's short-term behavior could provide a large benefit. Alternatively, the prediction quality with respect to point estimator and prediction interval could be improved through grouping and sorting of transitions based on equipment to increase the available database. Based on both domain knowledge and grouping algorithms provided by recent industrial implementations the results and computational effort could be improved.

Third, the prediction models can be expanded by regarding more algorithms, such as pattern recognition based approaches (Farahani et al. 2023) or novel time series models (Pham & Kuestenmacher, et al. 2023). Another frequently used approach is ensemble learning where several different model are trained and decision are derived through aggregation. Such ensemble learning could combine individual model benefits of good point estimators with sufficiently small prediction intervals which has been regarded on a general scale by Lakshminarayanan et al. (2017). Alternatively, the decision models can be simplified to improve speed and explainability for instance if simple decision rules are derived from complex machine learning models as introduced for reinforcement learning by Kuhnle et al. (2022). In the foresighted digital twin model several aspects are currently fit as the underlying algorithms are not available which results in a real-to-sim gap. Through inferring of algorithms based on observations for instance with black box based quantum computing could vastly reduce this real-to-sim and thus improve the predictions obtained during the foresight periods.

Fourth, the gate keeping decision which aims at reducing the number of time-constraint violations to improve the overall system performance could be augmented and combined to include scheduling decisions or be included into scheduling decision. For instance sequencing control could be derived that prioritizes time-constrained lots under certain circumstances to actively enforce time-constraint adherence. While the state-of-the-art research is far from regarding complex and large enough systems this approach is particularly promising if successfully implemented.

Last but not least, the approach should be applied to less complex job shops and other industries that exhibit time constraints such as in food processing, temperature based processes, gluing or in chemical treatments. One example is the control of galvano baths in electro-plating. This could allow other complex systems to reap the benefits of such a system and extened state of knowledge as a wider range of complex systems would be regarded.

7 Conclusion

Manufacturing under ever increasing complexities and stringent requirements increases the need for operational excellence which is fundamentally dependent on the possibility to invent ever improving intelligent production control. Intelligent production control for complex systems, such as controlling the gate keeping decision for time-constraints in complex job shops, relies on the availability of abundant and suitable data to exploit the full potential of self-improving artificial intelligence. While AI has become increasingly powerful complex job shops with time-constraints are still regarded with rigid models with strict assumptions and computational bounds that do not permit real-world applicability. Implementing machine learning and foresighted digital twins to provide a solution to time-constraints in complex job shops has not been sufficiently and successfully regarded.

To that end, this work overcomes this research deficit and presents a novel production control for gate keeping of time-constraints in complex job shops. The automated production control approach interacts with the real system and continuously updates its models. Beyond largely improving time-constraint adherence the currently manual effort required can be saved.

The presented comprehensive methodology to obtain probabilities for certain events in complex manufacturing systems through a combination of point and interval predictors is central to its success. Entangling the production system model with the real system through a digital twin that can provide foresight provides one solution. Additionally, single- and multivariate time series predictors are trained and their uncertainty evaluated to create the required prediction intervals. Solutions for the challenges of training, tuning and validating such models are implemented and presented. For prediction intervals the one-sided Winkler Loss as an evaluation metric is derived and novel methods to derive the prediction interval of various machine learning models are implemented. Based on the point estimator and prediction interval the known upper limit time-constraint can be evaluated resulting in the time-constraint adherence probability. The decision model is implemented upon this basis. Within the regarded validation time frames of the real-world semiconductor frontend wafer fab the proposed model outperforms the state-of-the-art and generates significant financial and environmental savings. The performance is robust over system changes over time. Thus, this research presents a strong contribution to intelligent production control for complex autonomous production systems. Not only current complex job shop operations managers in semiconductor manufacturing but all potentially complex production areas should evaluate the potential of this or related approaches to greatly improve operational performance and find a new sweet-spot of balancing complexity with flexibility and stringent product-inherent manufacturing constraints.

8 List of own publications

Behrendt & Ungen, et al. 2023

Behrendt, S.; Ungen, M.; Fisel, J.; Hung, K.-C.; May, M. C.; Leberle, U. & Lanza, G. (2023). “Improving Production System Flexibility and Changeability Through Software-Defined Manufacturing”. *Lecture Notes in Production Engineering*, pp. 705–716. DOI: 10.1007/978-3-031-18318-8_70.

Behrendt & Wurster, et al. 2023

Behrendt, S.; Wurster, M.; May, M. C. & Lanza, G. (2023). “Extended Production Planning of Reconfigurable Manufacturing Systems by Means of Simulation based Optimization”. *Proceedings of the Conference on Production Systems and Logistics: CPSL 2023*. Ed. by S. Thiede; G. Schuh & et al. (Santiago de Queretaro, Mexico, 28th Feb. - 3rd Mar. 2023): publish-Ing., pp. 210–219. DOI: 10.15488/13440.

Beiner & Kandler, et al. 2023

Beiner, S.; Kandler, M.; Richter, D.; May, M. C.; Kinkel, S. & Lanza, G. (2023). “Artificial Intelligence Implementation Strategy for Industrial Companies Using the AI Tool Box - A Morphology for Selecting Relevant AI Use Cases”. *Lecture Notes in Mechanical Engineering*, pp. 763–773. DOI: 10.1007/978-3-031-34821-1_83.

Brützel & Thiery, et al. 2022

Brützel, O.; Thiery, D.; May, M. & Lanza, G. (2022). “Optimization of a material flow control to increase energy efficiency in production”. *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb* 117.9, pp. 591–596. DOI: 10.1515/zwf-2022-1106.

Chen & Sampath, et al. 2023

Chen, T.; Sampath, V.; May, M. C.; Shan, S.; Jorg, O. J.; Aguilar Martín, J. J.; Stamer, F.; Fantoni, G.; Tosello, G. & Calaon, M. (2023). “Machine Learning in Manufacturing towards Industry 4.0: From ‘For Now’to ‘Four-Know’”. *Applied Sciences* 13.3, p. 1903. DOI: 10.3390/app13031903.

Frey & May, et al. 2022

Frey, A. M.; May, M. C. & Lanza, G. (2022). “Creation and validation of systems for product and process configuration based on data analysis”. *Production Engineering* 17.2, pp. 263–277. DOI: 10.1007/s11740-022-01176-1.

Hoffmann & Altenmüller, et al. 2021

Hoffmann, C.; Altenmüller, T.; May, M. C.; Kuhnle, A. & Lanza, G. (2021). “Simulative Dispatching Optimization of Maintenance Resources in a Semiconductor Use-Case Using Reinforcement Learning”. *Simulation in Produktion und Logistik 2021, Tagungsband 19. Fachtagung Simulation und Logistik 2021*. Ed. by J. Franke & P. Schuderer. (Erlangen,

Germany, 15th - 17th Sep. 2021): Cuvillier, pp. 357–366. DOI: https://d-nb.info/1240328044.

Hofmann & Liu, et al. 2022

Hofmann, C.; Liu, X.; May, M. & Lanza, G. (2022). “Hybrid Monte Carlo tree search based multi-objective scheduling”. *Production Engineering* 17.1, pp. 133–144. DOI: 10.1007/s11740-022-01152-9.

Kain & Nadimpalli, et al. 2020

Kain, M.; Nadimpalli, V. K.; Miqueo, A.; May, M. C.; Yagüe-Fabra, J. A.; Häfner, B.; Pedersen, D. B.; Calaon, M. & Tosello, G. (2020). “Metal additive manufacturing of multi-material dental strut implants”. *Proceedings of the 20th International Conference of the European Society for Precision Engineering and Nanotechnology, EUSPEN 2020*. Ed. by D. Caldwell; S. Carmignato & R. Leach. (Geneva, Switzerland, 8th - 12th Jun. 2020), pp. 175–176. DOI: 10.5445/IR/1000125403.

Kandler & Gabriel, et al. 2022

Kandler, M.; Gabriel, P.; Schröttle, V.; May, M. C. & Lanza, G. (2022). “Modular, Digital Shopfloor Management Model – A Maturity Assessment For A Human-Oriented Transformation Process”. *Proceedings of the Conference on Production Systems and Logistics: CPSL 2022*. Ed. by S. Thiede; G. Schuh & et al. (Vancouver, Canada, 17th - 20th May 2022): publish-Ing., pp. 642–651. DOI: 10.15488/12153.

Kandler & May, et al. 2022

Kandler, M.; May, M. C.; Kurtz, J.; Kuhnle, A. & Lanza, G. (2022). *Development of a Human-Centered Implementation Strategy for Industry 4.0 Exemplified by Digital Shopfloor Management*. Lecture Notes in Mechanical Engineering. Springer, pp. 738–745. DOI: 10.1007/978-3-030-90700-6_84.

Kandler & Dierolf, et al. 2022

Kandler, M.; Dierolf, L.; Bender, M.; Schäfer, L.; May, M. C. & Lanza, G. (2022). “Shopfloor Management Acceptance in Global Manufacturing”. *Procedia CIRP* 115, pp. 190–195. DOI: 10.1016/j.procir.2022.10.072.

Kapp & May, et al. 2020

Kapp, V.; May, M. C.; Lanza, G. & Wuest, T. (2020). “Pattern recognition in multivariate time series: Towards an automated event detection method for smart manufacturing systems”. *Journal of Manufacturing and Materials Processing* 4.3, p. 88. DOI: 10.3390/JMMP4030088.

Krahe & Marinov, et al. 2022

Krahe, C.; Marinov, M.; Schmutz, T.; Hermann, Y.; Bonny, M.; May, M. & Lanza, G. (2022). “AI based geometric similarity search supporting component reuse in engineering design”. *Procedia CIRP* 109, pp. 275–280. DOI: 10.1016/j.procir.2022.05.249.

Kuhn & May, et al. 2022

Kuhn, A. M.; May, M. C.; Liu, Y.; Kuhnle, A.; Tekouo, W. & Lanza, G. (2022). “Towards narrowing the reality gap in electromechanical systems: error modeling in virtual commissioning”. *Production Engineering* 17.3-4, pp. 535–545. DOI: 10.1007/s11740-022-01160-9.

Kuhnle & May, et al. 2022

Kuhnle, A.; May, M. C.; Schäfer, L. & Lanza, G. (2022). “Explainable reinforcement learning in production control of job shop manufacturing system”. *International Journal of Production Research* 60.19, pp. 5812–5834. DOI: 10.1080/00207543.2021.1972179.

Loffredo & May, et al. 2023

Loffredo, A.; May, M. C.; Schäfer, L.; Matta, A. & Lanza, G. (2023). “Reinforcement learning for energy-efficient control of parallel and identical machines”. *CIRP Journal of Manufacturing Science and Technology* 44, pp. 91–103. DOI: 10.1016/j.cirpj.2023.05.007.

May & Albers, et al. 2021

May, M. C.; Albers, A.; Fischer, M. D.; Mayerhofer, F.; Schäfer, L. & Lanza, G. (2021). “Queue Length Forecasting in Complex Manufacturing Job Shops”. *Forecasting* 3.2, pp. 322–338. DOI: 10.3390/forecast3020021.

May & Behnen, et al. 2021

May, M. C.; Behnen, L.; Holzer, A.; Kuhnle, A. & Lanza, G. (2021). “Multi-variate time-series for time constraint adherence prediction in complex job shops”. *Procedia CIRP* 103, pp. 55–60. DOI: 10.1016/j.procir.2021.10.008.

May & Fang, et al. 2022

May, M. C.; Fang, Z.; Eitel, M. B. M.; Stricker, N.; Ghoshdastidar, D. & Lanza, G. (2022). “Graph-based prediction of missing KPIs through optimization and random forests for KPI systems”. *Production Engineering* 17.2, pp. 211–222. DOI: 10.1007/s11740-022-01179-y.

May & Kiefer & Kuhnle & Lanza 2022

May, M. C.; Kiefer, L.; Kuhnle, A. & Lanza, G. (2022). “Ontology-Based Production Simulation with OntologySim”. *Applied Sciences* 12.3. DOI: 10.3390/app12031608.

May & Kiefer & Kuhnle & Stricker, et al. 2020

May, M. C.; Kiefer, L.; Kuhnle, A.; Stricker, N. & Lanza, G. (2020). “Decentralized Multi-Agent Production Control through Economic Model Bidding for Matrix Production Systems”. *Procedia CIRP* 96, pp. 3–8. DOI: 10.1016/j.procir.2021.01.043.

May & Kuhnle, et al. 2020

May, M. C.; Kuhnle, A. & Lanza, G. (2020). “Digitale Produktion und intelligente Steuerung – Integration von digitaler und realer Fertigung”. *wt Werkstattstechnik online* 110.4, pp. 255–260. DOI: 10.37544/1436-4980-2020-04-89.

May & Overbeck, et al. 2021
May, M. C.; Overbeck, L.; Wurster, M.; Kuhnle, A. & Lanza, G. (2021). “Foresighted digital twin for situational agent selection in production control”. *Procedia CIRP* 99, pp. 27–32. DOI: 10.1016/j.procir.2021.03.005.

May & Schmidt, et al. 2021
May, M. C.; Schmidt, S.; Kuhnle, A.; Stricker, N. & Lanza, G. (2021). “Product Generation Module: Automated Production Planning for optimized workload and increased efficiency in Matrix Production Systems”. *Procedia CIRP* 96, pp. 45–50. DOI: 10.1016/j.procir.2021.01.050.

May & Frenzer, et al. 2022
May, M. C.; Frenzer, M. & Lanza, G. (2022). “Teaching Machine Learning in Learning Factories with Industry 4.0 use-cases”. *Proceedings of the 12th Conference on Learning Factories (CLF 2022)*. Ed. by C. Ramsauer. (Singapore, Singapore, 11th - 13th Apr. 2022). DOI: 10.2139/ssrn.4071757.

May & Hermeler, et al. 2023
May, M. C.; Hermeler, S.; Mauch, E.; Dvorak, J.; Schäfer, L. & Lanza, G. (2023). “Reinforcement Learning for Improvement Measure Selection in Learning Factories”. *Proceedings of the 13th Conference on Learning Factories (CLF 2023)*. Ed. by C. Ramsauer & V. Hummel. (Reutlingen, Germany, 9th - 11th May. 2023). DOI: 10.2139/ssrn.4470426.

May & Kiefer & Frey, et al. 2023
May, M. C.; Kiefer, L.; Frey, A.; Duffie, N. A. & Lanza, G. (2023). “Solving sustainable aggregate production planning with model predictive control”. *CIRP Annals* 72 (1), pp. 421–424. DOI: 10.1016/j.cirp.2023.04.023.

May & Maucher, et al. 2021
May, M. C.; Maucher, S.; Holzer, A.; Kuhnle, A. & Lanza, G. (2021). “Data analytics for time constraint adherence prediction in a semiconductor manufacturing use-case”. *Procedia CIRP* 100, pp. 49–54. DOI: 10.1016/j.procir.2021.05.008.

May & Muriel, et al. 2018
May, M. C.; Muriel, A. & Nikolaidis, Y. (2018). “Analysis of Remanufacturing Suitability”. *Proceedings of the 4th International Conference of Green Supply Chain Management*. Ed. by Y. Nikolaidis. (Thessaloniki, Greece, 2nd - 4th Jul. 2018), pp. 229–240.

May & Neidhöfer, et al. 2022
May, M. C.; Neidhöfer, J.; Körner, T.; Schäfer, L. & Lanza, G. (2022). “Applying Natural Language Processing in Manufacturing”. *Procedia CIRP* 115, pp. 184–189. DOI: 10.1016/j.procir.2022.10.071.

May & Schäfer, et al. 2023
May, M. C.; Schäfer, L.; Frey, A.; Krahe, C. & Lanza, G. (2023). “Towards Product-Production-

CoDesign for the Production of the Future". *Procedia CIRP* 119, pp. 944–949. DOI: 10.1016/j.procir.2023.02.172.

Overbeck & Glaser, et al. 2023

Overbeck, L.; Glaser, V.; May, M. C. & Lanza, G. (2023). "Generalization of Reinforcement Learning Agents for Production Control". *Lecture Notes in Mechanical Engineering*, pp. 338–346. DOI: 10.1007/978-3-031-34821-1_37.

Overbeck & Hugues, et al. 2021

Overbeck, L.; Hugues, A.; May, M. C.; Kuhnle, A. & Lanza, G. (2021). "Reinforcement Learning Based Production Control of Semi-automated Manufacturing Systems". *Procedia CIRP* 103, pp. 170–175. DOI: 10.1016/j.procir.2021.10.027.

Overbeck & Rose, et al. 2022

Overbeck, L.; Rose, A.; May, M. & Lanza, G. (2022). "Utilization Concept for Digital Twins of Production Systems Integration into the Organization and Production Planning Processes". *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb* 117.4, pp. 244–248. DOI: 10.1515/zwf-2022-1035.

Schäfer & Frank, et al. 2022

Schäfer, L.; Frank, A.; May, M. C. & Lanza, G. (2022). "Automated Derivation of Optimal Production Sequences from Product Data". *Procedia CIRP* 107, pp. 469–474. DOI: 10.1016/j.procir.2022.05.010.

Schäfer & Treml, et al. 2023

Schäfer, L.; Treml, N.; May, M. C. & Lanza, G. (2023). "Classifying Parts using Feature Extraction and Similarity Assessment". *Procedia CIRP* 119, pp. 822–827. DOI: 10.1016/j.procir.2023.03.127.

Valet & Altenmüller, et al. 2022

Valet, A.; Altenmüller, T.; Waschneck, B.; May, M. C.; Kuhnle, A. & Lanza, G. (2022). "Opportunistic maintenance scheduling with deep reinforcement learning". *Journal of Manufacturing Systems* 64, pp. 518–534. DOI: 10.1016/j.jmsy.2022.07.016.

Wurster & Michel, et al. 2022

Wurster, M.; Michel, M.; May, M. C.; Kuhnle, A.; Stricker, N. & Lanza, G. (2022). "Modelling and condition-based control of a flexible and hybrid disassembly system with manual and autonomous workstations using reinforcement learning". *Journal of Intelligent Manufacturing* 33.2, pp. 575–591. DOI: 10.1007/s10845-021-01863-3.

9 References

Abdar & Pourpanah, et al. 2021

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R., et al. (2021). “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. *Information Fusion* 76, pp. 243–297. DOI: 10.1016/j.inffus.2021.05.008.

Abramovici & Göbel, et al. 2016

Abramovici, M.; Göbel, J. C. & Dang, H. B. (2016). “Semantic data management for the development and continuous reconfiguration of smart products and systems”. *CIRP Annals* 65 (1), pp. 185–188. DOI: 10.1016/j.cirp.2016.04.051.

Altenmüller & Stüker, et al. 2020

Altenmüller, T.; Stüker, T.; Waschneck, B.; Kuhnle, A. & Lanza, G. (2020). “Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints”. *Production Engineering* 14, pp. 319–328. DOI: 10.1007/s11740-020-00967-8.

An & Kim, et al. 2016

An, Y.-J.; Kim, Y.-D. & Choi, S.-W. (2016). “Minimizing makespan in a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times”. *Computers & Operations Research* 71, pp. 127–136. DOI: 10.1016/j.cor.2016.01.017.

Arima & Kobayashi, et al. 2015

Arima, S.; Kobayashi, A.; Wang, Y.-F.; Sakurai, K. & Monma, Y. (2015). “Optimization of Re-Entrant Hybrid Flows With Multiple Queue Time Constraints in Batch Processes of Semiconductor Manufacturing”. *IEEE Transactions on Semiconductor Manufacturing* 28.4, pp. 528–544. DOI: 10.1109/TSM.2015.2478281.

Arinez & Chang, et al. 2020

Arinez, J. F.; Chang, Q.; Gao, R. X.; Xu, C. & Zhang, J. (2020). “Artificial intelligence in advanced manufacturing: Current status and future outlook”. *Journal of Manufacturing Science and Engineering* 142 (11), p. 110804. DOI: 10.1115/1.4047855.

Askanazi & Diebold, et al. 2018

Askanazi, R.; Diebold, F. X.; Schorfheide, F. & Shin, M. (2018). “On the comparison of interval forecasts”. *Journal of Time Series Analysis* 39 (6), pp. 953–965. DOI: 10.1111/jtsa.12426.

Behrendt & Ungen, et al. 2023

Behrendt, S.; Ungen, M.; Fisel, J.; Hung, K.-C.; May, M. C.; Leberle, U. & Lanza, G. (2023). “Improving Production System Flexibility and Changeability Through Software-

Defined Manufacturing". *Lecture Notes in Production Engineering*, pp. 705–716. DOI: 10.1007/978-3-031-18318-8_70.

Bergstra & Yamins, et al. 2013

Bergstra, J.; Yamins, D. & Cox, D. (2013). "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures". *Proceedings of the 30th International Conference on Machine Learning*. Ed. by M. Littmann; S. Dasgupta & D. McAllester. Vol. 28. (Atlanta, USA, 16th - 21st Jun. 2013), pp. 115–123.

Bixby & Burda, et al. 2006

Bixby, R.; Burda, R. & Miller, D. (2006). "Short-Interval Detailed Production Scheduling in 300mm Semiconductor Manufacturing using Mixed Integer and Constraint Programming". *The 17th Annual SEMI/IEEE ASMC 2006 Conference*. Ed. by N. Govind & J. Tyminski. (Boston, USA, 22nd - 24th May. 2006), pp. 148–154. DOI: 10.1109/ASMC.2006.1638740.

Bock & Goppold, et al. 2018

Bock, S.; Goppold, J. & Weiß, M. (2018). "An improvement of the convergence proof of the ADAM-Optimizer". *arXiv preprint arXiv:1804.10587*. DOI: 10.48550/arXiv.1804.10587.

Burke & Gendreau, et al. 2013

Burke, E. K.; Gendreau, M.; Hyde, M.; Kendall, G.; Ochoa, G.; Özcan, E. & Qu, R. (2013). "Hyper-heuristics: A survey of the state of the art". *Journal of the Operational Research Society* 64.12, pp. 1695–1724. DOI: 10.1057/jors.2013.71.

Calvo & Yagüe-Fabra, et al. 2023

Calvo, R.; Yagüe-Fabra, J. A. & Tosello, G. (2023). "Advances in Sustainable and Digitalized Factories: Manufacturing, Measuring Technologies and Systems". *Applied Sciences* 13.9, p. 5570. DOI: 10.3390/app13095570.

Casella & Hwang, et al. 1993

Casella, G.; Hwang, J. G. & Robert, C. (1993). "A paradox in decision-theoretic interval estimation". *Statistica Sinica* 3.1, pp. 141–155.

Chakravorty & Nagarur 2020

Chakravorty, S. & Nagarur, N. N. (2020). "An Artificial Neural Network Based Algorithm For Real Time Dispatching Decisions". *31st Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC 2020)*. Ed. by B. Wood & V. Seshachalam. (Sarotoga Springs, USA, 24th - 26th Aug. 2020), pp. 1–5. DOI: 10.1109/ASMC49169.2020.9185213.

Chang & Chang 2012

Chang, C.-Y. & Chang, K.-H. (2012). "An integrated and improved dispatching approach to reduce cycle time of wet etch and furnace operations in semiconductor fabrication". *Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. Ed. by J. Barthes & J. Luo. (Wuhan, China, 23th - 25th May. 2012), pp. 734–741. DOI: 10.1109/CSCWD.2012.6221901.

Chatfield 2001

Chatfield, C. (2001). “Prediction Intervals for Time-Series Forecasting”. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Ed. by J. S. Armstrong. Boston, MA: Springer US, pp. 475–494. DOI: 10.1007/978-0-306-47630-3_21.

Chen & Yang 2006

Chen, J.-S. & Yang, J.-S. (2006). “Model formulations for the machine scheduling problem with limited waiting time constraints”. *Journal of Information and Optimization Sciences* 27.1, pp. 225–240. DOI: 10.1080/02522667.2006.10699688.

Chen & Sampath, et al. 2023

Chen, T.; Sampath, V.; May, M. C.; Shan, S.; Jorg, O. J.; Aguilar Martín, J. J.; Stamer, F.; Fantoni, G.; Tosello, G. & Calaon, M. (2023). “Machine Learning in Manufacturing towards Industry 4.0: From ‘For Now’to ‘Four-Know’”. *Applied Sciences* 13.3, p. 1903. DOI: 10.3390/app13031903.

Chen & Tosello, et al. 2022

Chen, T.; Tosello, G.; Werner, N. & Calaon, M. (2022). “Anomaly Detection in Float-Zone Crystal Growth of Silicon”. *Procedia CIRP* 107, pp. 1515–1519. DOI: 10.1016/j.procir.2022.05.184.

Chien & Chen 2007

Chien, C.-F. & Chen, C.-H. (2007). “A novel timetabling algorithm for a furnace process for semiconductor fabrication with constrained waiting and frequency-based setups”. *OR Spectrum* 29.3, pp. 391–419. DOI: 10.1007/s00291-006-0062-3.

Chien & Kuo, et al. 2020

Chien, C.-F.; Kuo, C.-J. & Yu, C.-M. (2020). “Tool allocation to smooth work-in-process for cycle time reduction and an empirical study”. *Annals of Operations Research* 290.1-2, pp. 1009–1033. DOI: 10.1007/s10479-018-3034-5.

Cho & Van Merriënboer, et al. 2014

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H. & Bengio, Y. (2014). “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. *arXiv preprint arXiv:1406.1078*. DOI: 10.48550/arXiv.1406.1078.

Cho & Park, et al. 2014

Cho, L.; Park, H. M.; Ryan, J. K.; Sharkey, T. C.; Jung, C. & Pabst, D. (2014). “Production scheduling with queue-time constraints: Alternative formulations”. *Proceedings of the IIE Annual Conference and Expo 2014*. Ed. by W. Cook. (Montreal, Canada, 31st May - 3rd Jun. 2014), pp. 282–291.

Cholette & Djurdjanovic 2014

Cholette, M. E. & Djurdjanovic, D. (2014). “Degradation modeling and monitoring of

machines using operation-specific hidden Markov models". *IIE Transactions* 46.10, pp. 1107–1123. DOI: 10.1080/0740817X.2014.905734.

Ciccullo & Pero, et al. 2014

Ciccullo, F.; Pero, M.; Pirovano, G. & Sianesi, A. (2014). "Scheduling batches with time constraints in a job shop system: developing two approaches for semiconductor industry". *XIX Summer School "Francesco Turco" - Industrial Mechanical Plants*. Ed. by M. Bevilacqua. (Senigalli, Italy, 9th - 12th Sep. 2014), p. 12.

Cortes & Jackel, et al. 1994

Cortes, C.; Jackel, L. D. & Chiang, W.-P. (1994). "Limits on learning machine accuracy imposed by data quality". *Advances in Neural Information Processing Systems* 7.

Czumanski & Lödding 2016

Czumanski, T. & Lödding, H. (2016). "State-based analysis of labour productivity". *International Journal of Production Research* 54.10, pp. 2934–2950. DOI: 10.1080/00207543.2015.1137372.

Dodge 2008

Dodge, Y. (2008). *The concise encyclopedia of statistics*. 1st ed. New York, NY: Springer. DOI: 10.1007/978-0-387-32833-1.

Domingos 2012

Domingos, P. (2012). "A few useful things to know about machine learning". *Communications of the ACM* 55.10, pp. 78–87.

Dunke & Nickel 2016

Dunke, F. & Nickel, S. (2016). "A general modeling approach to online optimization with lookahead". *Omega* 63, pp. 134–153. DOI: 10.1016/j.omega.2015.10.009.

Eversheim & Schuh 2013

Eversheim, W. & Schuh, G. (2013). *Produktion und Management 3: Gestaltung von Produktionssystemen*. 52nd ed. New York, NY: Springer. DOI: 10.1007/978-3-642-58399-5.

Farahani & McCormick, et al. 2023

Farahani, M. A.; McCormick, M.; Gianinny, R.; Hudacheck, F.; Harik, R.; Liu, Z. & Wuest, T. (2023). "Time-series pattern recognition in Smart Manufacturing Systems: A literature review and ontology". *Journal of Manufacturing Systems* 69, pp. 208–241. DOI: 10.1016/j.jmsy.2023.05.025.

Filion 2015

Filion, G. J. (2015). "The signed Kolmogorov-Smirnov test: why it should not be used". *Gigascience* 4 (1), pp. 13742–015. DOI: 10.1186/s13742-015-0048-7.

Gabel & Riedmiller 2012

Gabel, T. & Riedmiller, M. (2012). "Distributed policy search reinforcement learning for

job-shop scheduling tasks". *International Journal of Production Research* 50.1, pp. 41–61. DOI: 10.1080/00207543.2011.571443.

Gal & Ghahramani 2016

Gal, Y. & Ghahramani, Z. (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". *Proceedings of the 33rd International Conference on Machine Learning*. Ed. by J. Langford; N. Balcan & K. Weinberger. Vol. 48. (New York City, USA, 19th - 24th Jun. 2016), pp. 1050–1059.

Gneiting & Raftery 2007

Gneiting, T. & Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American statistical Association* 102.477, pp. 359–378. DOI: 10.1198/016214506000001437.

Goodfellow & Bengio, et al. 2016

Goodfellow, I.; Bengio, Y. & Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.

Graff & Hanasusanto, et al. 2023

Graff, N.; Hanasusanto, G. A. & Djurdjanović, D. (2023). "Robust control of maximum photolithography overlay error in a pattern layer". *CIRP Annals* 72 (1), pp. 429–432. DOI: 10.1016/j.cirp.2023.03.015.

Gruber 1995

Gruber, T. R. (1995). "Toward principles for the design of ontologies used for knowledge sharing?" *International Journal of Human-Computer Studies* 43.5-6, pp. 907–928. DOI: 10.1006/ijhc.1995.1081.

Gu & Koren 2018

Gu, X. & Koren, Y. (2018). "Manufacturing system architecture for cost-effective mass-individualization". *Manufacturing letters* 16, pp. 44–48. DOI: 10.1016/j.mfglet.2018.04.002.

Gudivada & Apon, et al. 2017

Gudivada, V.; Apon, A. & Ding, J. (2017). "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations". *International Journal on Advances in Software* 10.1, pp. 1–20.

Ham & Lee, et al. 2011

Ham, M.; Lee, Y. H. & An, J. (2011). "IP-Based Real-Time Dispatching for Two-Machine Batching Problem With Time Window Constraints". *IEEE Transactions on Automation Science and Engineering* 8.3, pp. 589–597. DOI: 10.1109/TASE.2010.2098867.

Ham & Raiford, et al. 2006

Ham, M.; Raiford, M.; Dillard, F.; Risner, W.; Knisely, M.; Harrington, J.; Murtha, T. & Park, H. (2006). "Dynamic wet-furnace dispatching/scheduling in wafer fab". *The 17th Annual*

SEMI/IEEE ASMC 2006 Conference. Ed. by N. Govind & J. Tyminski. IEEE. (Boston, USA, 22nd - 24th May. 2006), pp. 144–147. DOI: 10.1109/ASMC.2006.1638739.

Handl & Kuhlenkasper 2018

Handl, A. & Kuhlenkasper, T. (2018). *Einführung in die Statistik: Theorie und Praxis mit R*. Heidelberg: Springer-Verlag. DOI: 10.1007/978-3-662-56440-0.

Helpman 1981

Helpman, E. (1981). "International trade in the presence of product differentiation, economies of scale and monopolistic competition: A Chamberlin-Heckscher-Ohlin approach". *Journal of International Economics* 11.3, pp. 305–340. DOI: 10.1016/0022-1996(81)90001-5.

Hochreiter & Schmidhuber 1997

Hochreiter, S. & Schmidhuber, J. (Dec. 1997). "Long Short-term Memory". *Neural Computation* 9, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

Hoffmann & Altenmüller, et al. 2021

Hoffmann, C.; Altenmüller, T.; May, M. C.; Kuhnle, A. & Lanza, G. (2021). "Simulative Dispatching Optimization of Maintenance Resources in a Semiconductor Use-Case Using Reinforcement Learning". *Simulation in Produktion und Logistik 2021, Tagungsband 19. Fachtagung Simulation und Logistik 2021*. Ed. by J. Franke & P. Schuderer. (Erlangen, Germany, 15th - 17th Sep. 2021): Cuvillier, pp. 357–366. DOI: https://d-nb.info/1240328044.

Hogan & Blomqvist, et al. 2021

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; Staab, S., et al. (2021). "Knowledge graphs". *ACM Computing Surveys* 54.4, pp. 1–37. DOI: 10.1145/3447772.

Hong & Chien, et al. 2023

Hong, T.-Y.; Chien, C.-F. & Chen, H.-P. (2023). "UNISON framework of system dynamics-based technology acquisition decision for semiconductor manufacturing and an empirical study". *Computers & Industrial Engineering*, p. 109012. DOI: 10.1016/j.cie.2023.109012.

Huang & Ke, et al. 2011

Huang, W.-Y.; Ke, L. & Shen, T. (2011). "Quantify equipment capacity impacts induced by maximum waiting time constraint through simulation". *2011 e-Manufacturing & Design Collaboration Symposium & International Symposium on Semiconductor Manufacturing (eMDC & ISSM)*. Ed. by C.-J. Lu; K.-C. Chen & T. Yang. (Hsinchu, Taiwan, 5th - 6th Sep. 2011), pp. 1–3.

Ingenieure 1996

Ingenieure, V. D. (1996). *VDI 3633 Simulation von Logistik-, Materialfluß- und Produktionssytemen - VDI 3633 Entwurf | Begriffsdefinitionen*. Düsseldorf: Beuth.

Irani & Cheng, et al. 1993
Irani, K. B.; Cheng, J.; Fayyad, U. M. & Qian, Z. (1993). “Applying machine learning to semiconductor manufacturing”. *IEEE Expert* 8.1, pp. 41–47. DOI: 10.1109/64.193054.

Jia & Jiang, et al. 2013
Jia, W.; Jiang, Z. & Li, Y. (2013). “Closed loop control-based real-time dispatching heuristic on parallel batch machines with incompatible job families and dynamic arrivals”. *International Journal of Production Research* 51.15, pp. 4570–4584. DOI: 10.1080/00207543.2013.774505.

Jiang & Deng 2023
Jiang, X. & Deng, X. (2023). “Knowledge reverse distillation based confidence calibration for deep neural networks”. *Neural Processing Letters* 55.1, pp. 345–360. DOI: 10.1007/s11063-022-10885-8.

Jørgensen & Sjoeberg 2003
Jørgensen, M. & Sjoeberg, D. I. K. (2003). “An effort prediction interval approach based on the empirical distribution of previous estimation accuracy”. *Information and Software Technology* 45.3, pp. 123–136. DOI: 10.1016/S0950-5849(02)00188-X.

Jung & Pabst, et al. 2013
Jung, C.; Pabst, D.; Ham, M.; Stehli, M. & Rothe, M. (2013). “An Effective Problem Decomposition Method for Scheduling of Diffusion Processes Based on Mixed Integer Linear Programming”. *24th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. Ed. by R. Dover. Vol. 24. (Piscataway, NJ, USA, 14th - 16th May. 2013), pp. 35–40. DOI: 10.1109/ASMC.2013.6552771.

Jung & Pabst, et al. 2014
Jung, C.; Pabst, D.; Ham, M.; Stehli, M. & Rothe, M. (2014). “An Effective Problem Decomposition Method for Scheduling of Diffusion Processes Based on Mixed Integer Linear Programming”. *IEEE Transactions on Semiconductor Manufacturing* 27.3, pp. 357–363. DOI: 10.1109/TSM.2014.2337310.

Jurisica & Mylopoulos, et al. 2004
Jurisica, I.; Mylopoulos, J. & Yu, E. (2004). “Ontologies for knowledge management: an information systems perspective”. *Knowledge and Information Systems* 6.4, pp. 380–401. DOI: 10.1007/s10115-003-0135-4.

Kabir & Khosravi, et al. 2018
Kabir, H. D.; Khosravi, A.; Hosen, M. A. & Nahavandi, S. (2018). “Neural network-based uncertainty quantification: A survey of methodologies and applications”. *IEEE Access* 6, pp. 36218–36234. DOI: 10.1109/ACCESS.2018.2836917.

Kao & Zhan, et al. 2011
Kao, Y.-T.; Zhan, S.-C.; Chang, S.-C.; Ho, J.-H.; Wang, P.; Luh, P. B.; Wang, S.; Wang, F.

& Chang, J. (2011). “Near optimal furnace tool allocation with batching and waiting time constraints”. *2011 IEEE International Conference on Automation Science and Engineering (CASE)*. Ed. by M. Wang. (Trieste, Italy, 24th - 27th Aug. 2011), pp. 108–113. DOI: 10.1109/CASE.2011.6042507.

Karsu & Morton 2015

Karsu, Ö. & Morton, A. (2015). “Inequity averse optimization in operational research”. *European Journal of Operational Research* 245.2, pp. 343–359. DOI: 10.1016/j.ejor.2015.02.035.

Khosravi & Nahavandi, et al. 2011

Khosravi, A.; Nahavandi, S.; Creighton, D. & Atiya, A. F. (2011). “Comprehensive review of neural network-based prediction intervals and new advances”. *IEEE Transactions on Neural Networks* 22.9, pp. 1341–1356. DOI: 10.1109/TNN.2011.2162110.

Kilby 2000

Kilby, J. S. (2000). “The integrated circuit’s early history”. *Proceedings of the IEEE* 88.1, pp. 109–111. DOI: 10.1109/5.811607.

Kim & Lee 2017

Kim, H.-J. & Lee, J.-H. (2017). “A Branch and Bound Algorithm for Three-Machine Flow Shop with Overlapping Waiting Time Constraints”. *IFAC-PapersOnLine* 50.1, pp. 1101–1105. DOI: 10.1016/j.ifacol.2017.08.391.

Kitamura & Mori, et al. 2006

Kitamura, S.; Mori, K. & Ono, A. (2006). “Capacity Planning Method for Semiconductor Fab with Time Constraints between Operations”. *2006 SICE-ICASE International Joint Conference*. Ed. by A. Nagashima. (Busan, Korea, 18th - 21st Aug. 2006), pp. 1100–1103. DOI: 10.1109/SICE.2006.315820.

Klemmt & Horn, et al. 2008

Klemmt, A.; Horn, S.; Weigert, G. & Hielscher, T. (2008). “Simulations-based and solver-based optimization approaches for batch processes in semiconductor manufacturing”. *2008 Winter Simulation Conference*. Ed. by T. Jefferson & J. Fowler. (Miami, FL, USA, 7th - 10th Dec. 2008), pp. 2041–2049. DOI: 10.1109/WSC.2008.4736300.

Klemmt & Mönch 2012

Klemmt, A. & Mönch, L. (2012). “Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing”. *Proceedings of the 2012 Winter Simulation Conference (WSC)*. Ed. by O. Rose & A. Uhrmacher. (Berlin, Germany, 9th - 12th Dec. 2012), pp. 1–10. DOI: 10.1109/WSC.2012.6465235.

Knublauch & Oberle, et al. 2006

Knublauch, H.; Oberle, D.; Tetlow, P.; Wallace, E.; Pan, J. & Uschold, M. (2006). “A semantic web primer for object-oriented software developers”. *W3C working group note*.

Kobayashi & Kuno, et al. 2013
Kobayashi, A.; Kuno, T. & Arima, S. (2013). “Re-entrant flow control in Q-time constraints processes for actual applications”. *2013 e-Manufacturing & Design Collaboration Symposium (eMDC)*. Ed. by L. Tuung. (Hsinchu, Taiwan, 6th Sep. 2013), pp. 1–4. DOI: 10.1109/eMDC.2013.6756052.

Koller & Friedman, et al. 2007
Koller, D.; Friedman, N.; Džeroski, S.; Sutton, C.; McCallum, A.; Pfeffer, A.; Abbeel, P.; Wong, M.-F.; Heckerman, D.; Meek, C., et al. (2007). *Introduction to statistical relational learning*. Cambridge, MA: MIT press.

Kopp & Hassoun, et al. 2020
Kopp, D.; Hassoun, M.; Kalir, A. & Mönch, L. (2020). “Integrating Critical Queue Time Constraints Into SMT2020 Simulation Models”. *Proceedings of the 2020 Winter Simulation Conference (WSC)*. Ed. by R. Thiesing. (Vienna, Austria, 14th - 18h Dec. 2020), pp. 1813–1824. DOI: 10.1109/WSC48552.2020.9383889.

Koren & Heisel, et al. 1999
Koren, Y.; Heisel, U.; Jovane, F.; Moriwaki, T.; Pritschow, G.; Ulsoy, G. & Van Brussel, H. (1999). “Reconfigurable manufacturing systems”. *CIRP Annals* 48.2, pp. 527–540. DOI: 10.1016/S0007-8506(07)63232-6.

Krahe & Marinov, et al. 2022
Krahe, C.; Marinov, M.; Schmutz, T.; Hermann, Y.; Bonny, M.; May, M. & Lanza, G. (2022). “AI based geometric similarity search supporting component reuse in engineering design”. *Procedia CIRP* 109, pp. 275–280. DOI: 10.1016/j.procir.2022.05.249.

Kuhnle & May, et al. 2022
Kuhnle, A.; May, M. C.; Schäfer, L. & Lanza, G. (2022). “Explainable reinforcement learning in production control of job shop manufacturing system”. *International Journal of Production Research* 60.19, pp. 5812–5834. DOI: 10.1080/00207543.2021.1972179.

Kunst & Wagner 2020
Kunst, R. M. & Wagner, M. (2020). “Econometric forecasting: editors’ introduction”. *Empirical Economics* 58. DOI: 10.1007/s00181-019-01820-3.

Kuo & Chien, et al. 2011
Kuo, C.-J.; Chien, C.-F. & Chen, J.-D. (2011). “Manufacturing Intelligence to Exploit the Value of Production and Tool Data to Reduce Cycle Time”. *IEEE Transactions on Automation Science and Engineering* 8.1, pp. 103–111. DOI: 10.1109/TASE.2010.2040999.

L’Heureux & Grolinger, et al. 2017
L’Heureux, A.; Grolinger, K.; Elyamany, H. F. & Capretz, M. A. (2017). “Machine learning with big data: Challenges and approaches”. *IEEE Access* 5, pp. 7776–7797. DOI: 10.1109/ACCESS.2017.2696365.

Lakshminarayanan & Pritzel, et al. 2017
Lakshminarayanan, B.; Pritzel, A. & Blundell, C. (2017). “Simple and scalable predictive uncertainty estimation using deep ensembles”. Ed. by I. Guyon; U. Von Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan & R. Garnett. Vol. 30. (Long Beach, CA, USA, 4th - 9th Dec. 2017).

Lamy 2016
Lamy, J.-B. (2016). “Ontology-oriented programming for biomedical informatics”. *Transforming Healthcare with the Internet of Things*. Ed. by J. Hodfijk; B. Seroussi & C. Lovis. Amsterdam, Netherlands: IOS Press, pp. 64–68.

Lanza & Ferdows, et al. 2019
Lanza, G.; Ferdows, K.; Kara, S.; Mourtzis, D.; Schuh, G.; Váncza, J.; Wang, L. & Wiendahl, H.-P. (2019). “Global production networks: Design and operation”. *CIRP Annals* 68.2, pp. 823–841. DOI: 10.1016/j.cirp.2019.05.008.

Lasi & Fettke, et al. 2014
Lasi, H.; Fettke, P.; Kemper, H.-G.; Feld, T. & Hoffmann, M. (2014). “Industry 4.0”. *Business & Information Systems Engineering* 6, pp. 239–242. DOI: 10.1007/s12599-014-0334-4.

Lee 2020
Lee, J.-Y. (2020). “A genetic algorithm for a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times”. *Mathematical Problems in Engineering* 2020. DOI: 10.1155/2020/8833645.

Lee & Li 2017
Lee, J.-H. & Li, J. (2017). “Performance Evaluation of Bernoulli Serial Lines with Waiting Time Constraints”. *IFAC-PapersOnLine* 50.1, pp. 1087–1092. DOI: 10.1016/j.ifacol.2017.08.387.

Lee & Chen, et al. 2005
Lee, Y.-Y.; Chen, C. T. & Wu, C. (2005). “Reaction chain of process queue time quality control”. *ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005*. Ed. by J. Doran. (Piscataway, NJ, USA, 13th - 15th Sep. 2005): IEEE, pp. 47–50. DOI: 10.1109/ISSM.2005.1513293.

Li & Li, et al. 2012
Li, L.; Li, Y. F. & Sun, Z. J. (2012). “Dispatching rule considering time-constraints on processes for semiconductor wafer fabrication facility”. *2012 IEEE International Conference on Automation Science and Engineering (CASE)*. Ed. by N. Y. Chong. (Seoul, Korea, 20th - 24th Aug. 2012), pp. 407–412. DOI: 10.1109/CoASE.2012.6386370.

Lima & Borodin, et al. 2017a
Lima, A.; Borodin, V.; Dauzère-Pérès, S. & Vialletelle, P. (2017a). “A decision support system for managing line stops of time constraint tunnels: FA, IE”. *2017 28th Annual*

SEMI Advanced Semiconductor Manufacturing Conference (ASMC). Ed. by R. Dover & D. LeCunff. (Saratoga Springs, NY, USA, 15th - 18th May. 2017), pp. 309–314. DOI: 10.1109/ASMC.2017.7969250.

Lima & Borodin, et al. 2017b

Lima, A.; Borodin, V.; Dauzère-Pérès, S. & Vialletelle, P. (2017b). "Analyzing different dispatching policies for probability estimation in time constraint tunnels in semiconductor manufacturing". *2017 Winter Simulation Conference (WSC)*. Ed. by L. H. Lee & B. Nelson. (Las Vegas, NV, USA, 3th - 6th Dec. 2017), pp. 3543–3554. DOI: 10.1109/WSC.2017.8248068.

Lima & Borodin, et al. 2019

Lima, A.; Borodin, V.; Dauzère-Pérès, S. & Vialletelle, P. (2019). "Sampling-based release control of multiple lots in time constraint tunnels". *Computers in Industry* 110, pp. 3–11. DOI: 10.1016/j.compind.2019.04.014.

Lima & Borodin, et al. 2021

Lima, A.; Borodin, V.; Dauzère-Pérès, S. & Vialletelle, P. (2021). "A sampling-based approach for managing lot release in time constraint tunnels in semiconductor manufacturing". *International Journal of Production Research* 59.3, pp. 860–884. DOI: 10.1080/00207543.2020.1711984.

Little 2011

Little, J. D. (2011). "OR FORUM—Little's Law as viewed on its 50th anniversary". *Operations Research* 59.3, pp. 536–549. DOI: 10.1287/opre.1110.0940.

Liu & Jiang, et al. 2020

Liu, C.; Jiang, P. & Jiang, W. (2020). "Web-based digital twin modeling and remote control of cyber-physical production systems". *Robotics and Computer-integrated Manufacturing* 64, p. 101956. DOI: 10.1016/j.rcim.2020.101956.

Maleck & Eckert 2017

Maleck, C. & Eckert, T. (2017). "A comparison of control methods for production areas with time constraints and tool interruptions in semiconductor manufacturing". *2017 40th International Spring Seminar on Electronics Technology (ISSE)*. Ed. by A. Hamacek & N. Hinov. (Sofia, Bulgaria, 10th - 14th May. 2017), pp. 1–6. DOI: 10.1109/ISSE.2017.8000944.

Maleck & Nieke & Bock & Pabst & Schulze, et al. 2019

Maleck, C.; Nieke, G.; Bock, K.; Pabst, D.; Schulze, M. & Stehli, M. (2019). "A Robust Multi-Stage Scheduling Approach for Semiconductor Manufacturing Production Areas with Time Contraints". *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. Ed. by A. Jain & F. Levitov. (Saratoga Springs, NY, USA, 6th - 9th May. 2019), pp. 1–6. DOI: 10.1109/ASMC.2019.8791779.

Maleck & Nieke & Bock & Pabst & Stehli 2018

Maleck, C.; Nieke, G.; Bock, K.; Pabst, D. & Stehli, M. (2018). “A comparison of an CP and MIP approach for scheduling jobs in production areas with time constraints and uncertainties”. *2018 Winter Simulation Conference (WSC)*. Ed. by S. Breor; P. Frazier & R. Cheng. (Gothenburg, Sweden, 9th - 12th Dec. 2018), pp. 3526–3537. DOI: 10.1109/WSC.2018.8632404.

Maleck & Weigert, et al. 2017

Maleck, C.; Weigert, G.; Pabst, D. & Stehli, M. (2017). “Robustness analysis of an MIP for production areas with time constraints and tool interruptions in semiconductor manufacturing”. *2017 Winter Simulation Conference (WSC)*. Ed. by L. H. Lee & B. Nelson. (Las Vegas, NV, USA, 3rd - 6th Dec. 2017), pp. 3714–3725. DOI: 10.1109/WSC.2017.8248084.

Mason & Kurz, et al. 2007

Mason, S. J.; Kurz, M. E.; Pfund, M. E.; Fowler, J. W. & Pohl, L. M. (2007). “Multi-objective semiconductor manufacturing scheduling: A random keys implementation of NSGA-II”. Ed. by G. Kendall; K. C. Tan; E. Burke & S. Smith. (Honolulu, HI, USA, 1st - 5th Apr. 2007), pp. 159–164. DOI: 10.1109/SCIS.2007.367684.

Maté & Trujillo, et al. 2012

Maté, A.; Trujillo, J. & Mylopoulos, J. (2012). “Conceptualizing and specifying key performance indicators in business strategy models”. *Conceptual Modeling. ER 2012. Lecture Notes in Computer Science*. Ed. by P. Atzeni; D. Cheung & S. Ram. Vol. 7532. Heidelberg: Springer, pp. 282–291. DOI: 10.1007/978-3-642-34002-4_22.

May & Albers, et al. 2021

May, M. C.; Albers, A.; Fischer, M. D.; Mayerhofer, F.; Schäfer, L. & Lanza, G. (2021). “Queue Length Forecasting in Complex Manufacturing Job Shops”. *Forecasting* 3.2, pp. 322–338. DOI: 10.3390/forecast3020021.

May & Behnen, et al. 2021

May, M. C.; Behnen, L.; Holzer, A.; Kuhnle, A. & Lanza, G. (2021). “Multi-variate time-series for time constraint adherence prediction in complex job shops”. *Procedia CIRP* 103, pp. 55–60. DOI: 10.1016/j.procir.2021.10.008.

May & Fang, et al. 2022

May, M. C.; Fang, Z.; Eitel, M. B. M.; Stricker, N.; Ghoshdastidar, D. & Lanza, G. (2022). “Graph-based prediction of missing KPIs through optimization and random forests for KPI systems”. *Production Engineering* 17.2, pp. 211–222. DOI: 10.1007/s11740-022-01179-y.

May & Kiefer & Kuhnle & Lanza 2022
May, M. C.; Kiefer, L.; Kuhnle, A. & Lanza, G. (2022). “Ontology-Based Production Simulation with OntologySim”. *Applied Sciences* 12.3. DOI: 10.3390/app12031608.

May & Kuhnle, et al. 2020
May, M. C.; Kuhnle, A. & Lanza, G. (2020). “Digitale Produktion und intelligente Steuerung – Integration von digitaler und realer Fertigung”. *wt Werkstattstechnik online* 110.4, pp. 255–260. DOI: 10.37544/1436-4980-2020-04-89.

May & Overbeck, et al. 2021
May, M. C.; Overbeck, L.; Wurster, M.; Kuhnle, A. & Lanza, G. (2021). “Foresighted digital twin for situational agent selection in production control”. *Procedia CIRP* 99, pp. 27–32. DOI: 10.1016/j.procir.2021.03.005.

May & Schmidt, et al. 2021
May, M. C.; Schmidt, S.; Kuhnle, A.; Stricker, N. & Lanza, G. (2021). “Product Generation Module: Automated Production Planning for optimized workload and increased efficiency in Matrix Production Systems”. *Procedia CIRP* 96, pp. 45–50. DOI: 10.1016/j.procir.2021.01.050.

May & Kiefer & Frey, et al. 2023
May, M. C.; Kiefer, L.; Frey, A.; Duffie, N. A. & Lanza, G. (2023). “Solving sustainable aggregate production planning with model predictive control”. *CIRP Annals* 72 (1), pp. 421–424. DOI: 10.1016/j.cirp.2023.04.023.

May & Kiefer & Kuhnle & Stricker, et al. 2021
May, M. C.; Kiefer, L.; Kuhnle, A.; Stricker, N. & Lanza, G. (2021). “Decentralized multi-agent production control through economic model bidding for matrix production systems”. *Procedia CIRP* 96, pp. 3–8.

May & Maucher, et al. 2021
May, M. C.; Maucher, S.; Holzer, A.; Kuhnle, A. & Lanza, G. (2021). “Data analytics for time constraint adherence prediction in a semiconductor manufacturing use-case”. *Procedia CIRP* 100, pp. 49–54. DOI: 10.1016/j.procir.2021.05.008.

Mazzola & Kapahnke, et al. 2016
Mazzola, L.; Kapahnke, P.; Vujic, M. & Klusch, M. (2016). “CDM-Core: A Manufacturing Domain Ontology in OWL2 for Production and Maintenance.” *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 2: KEOD*. Ed. by A. Fred; J. Dietz & D. Aveiro. (Porto, Portugal, 9th - 11th Nov. 2016), pp. 136–143.

McGuinness & Van Harmelen, et al. 2004
McGuinness, D. L.; Van Harmelen, F., et al. (2004). “OWL web ontology language overview”.

W3C recommendation 10.10. [last accessed 15.10.2023]. URL: https://www.w3.org/TR/owl-features/.

Min & Lu, et al. 2019

Min, Q.; Lu, Y.; Liu, Z.; Su, C. & Wang, B. (2019). "Machine learning based digital twin framework for production optimization in petrochemical industry". *International Journal of Information Management* 49, pp. 502–519. DOI: 10.1016/j.ijinfomgt.2019.05.020.

Mitchell 2006

Mitchell, T. M. (2006). *The discipline of machine learning*. Vol. 9. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning.

Mönch & Fowler & Dauzère-Pérès, et al. 2009

Mönch, L.; Fowler, J. W.; Dauzère-Pérès, S.; Mason, S. J. & Rose, O. (2009). "Scheduling semiconductor manufacturing operations: Problems, solution techniques, and future challenges". *4th Multidisciplinary International Conference on Scheduling: Theory & Applications (MISTA)*. Ed. by E. Pesch. (Dublin, Ireland, 10th - 12th Aug. 2009). DOI: 10.1007/s10951-010-0222-9.

Mönch & Fowler & Dauzère-Pérès, et al. 2011

Mönch, L.; Fowler, J. W.; Dauzère-Pérès, S.; Mason, S. J. & Rose, O. (2011). "A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations". *Journal of Scheduling* 14.6, pp. 583–599. DOI: 10.1007/s10951-010-0222-9.

Mönch & Fowler & Mason 2013

Mönch, L.; Fowler, J. W. & Mason, S. J. (2013). *Production planning and control for semiconductor wafer fabrication facilities: Modeling, analysis, and systems*. 52nd ed. Operations Research / Computer Science Interfaces Series. New York, NY: Springer. DOI: 10.1007/978-1-4614-4472-5.

Mönch & Stehli 2003

Mönch, L. & Stehli, M. (2003). "An ontology for production control of semiconductor manufacturing processes". *Multiagent System Technologies. MATES 2003. Lecture Notes in Computer Science*. Ed. by M. Schillo; M. Klusch; J. Müller & H. Tianfield. Vol. 2831. Springer, pp. 156–167. DOI: 10.1007/978-3-540-39869-1_14.

Monostori & Csáji, et al. 2004

Monostori, L.; Csáji, B. C. & Kádár, B. (2004). "Adaptation and learning in distributed production control". *CIRP Annals* 53.1, pp. 349–352. DOI: 10.1016/S0007-8506(07)60714-8.

Morcos & Barrett, et al. 2018

Morcos, A. S.; Barrett, D. G.; Rabinowitz, N. C. & Botvinick, M. (2018). "On the importance

of single directions for generalization". *arXiv preprint arXiv:1803.06959*. DOI: 10.48550/arXiv.1803.06959.

Murty 1994

Murty, K. G. (1994). *Operations research: deterministic optimization models*. 52nd ed. New Jersey, US: Prentice-Hall, Inc. DOI: 10.5555/185059.

Nattaf & Dauzère-Pérès, et al. 2019

Nattaf, M.; Dauzère-Pérès, S.; Yugma, C. & Wu, C.-H. (2019). "Parallel machine scheduling with time constraints on machine qualifications". *Computers & Operations Research* 107, pp. 61–76. DOI: 10.1016/j.cor.2019.03.004.

Negri & Fumagalli, et al. 2017

Negri, E.; Fumagalli, L. & Macchi, M. (2017). "A review of the roles of digital twin in CPS-based production systems". *Procedia Manufacturing* 11, pp. 939–948. DOI: 10.1016/j.promfg.2017.07.198.

Nickel & Stein, et al. 2014

Nickel, S.; Stein, O. & Waldmann, K.-H. (2014). *Operations Research*. 2nd ed. Berlin, Germany: Springer. DOI: 10.1007/978-3-642-54368-5.

Ono & Kitamura, et al. 2006

Ono, A.; Kitamura, S. & Mori, K. (2006). "Risk Based Capacity Planning Method for Semiconductor Fab with Queue Time Constraints". *2006 IEEE International Symposium on Semiconductor Manufacturing*. Ed. by H. Sasaki & T. Ohmi. (Tokyo, Japan, 25th - 27st Sep. 2006), pp. 49–52. DOI: 10.1109/ISSM.2006.4493020.

Overbeck & Hugues, et al. 2021

Overbeck, L.; Hugues, A.; May, M. C.; Kuhnle, A. & Lanza, G. (2021). "Reinforcement Learning Based Production Control of Semi-automated Manufacturing Systems". *Procedia CIRP* 103, pp. 170–175. DOI: 10.1016/j.procir.2021.10.027.

Overbeck & Rose, et al. 2022

Overbeck, L.; Rose, A.; May, M. & Lanza, G. (2022). "Utilization Concept for Digital Twins of Production Systems Integration into the Organization and Production Planning Processes". *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb* 117.4, pp. 244–248. DOI: 10.1515/zwf-2022-1035.

Pappert & Zhang, et al. 2016

Pappert, F. S.; Zhang, T.; Rose, O.; Suhrke, F.; Mager, J. & Frey, T. (2016). "Impact of time bound constraints and batching on metallization in an opto-semiconductor fab". *Proceedings of the 2016 Winter Simulation Conference (WSC)*. Ed. by P. Frazier; T. Roeder & E. Zhou. (Washington D.C., USA, 11th - 14th Dec. 2016), pp. 2947–2957. DOI: 10.1109/WSC.2016.7822329.

Paternina-Arboleda & Das 2001

Paternina-Arboleda, C. D. & Das, T. K. (2001). “Intelligent dynamic control policies for serial production lines”. *IIE Transactions* 33.1, pp. 65–77. DOI: 10.1023/A:1007641824604.

Pe 2018

Pe, H. G. (2018). *Semiconductor Manufacturing Handbook*. 2nd ed. New York, NY: McGraw-Hill Education. DOI: 10.1036/978-1-25-958312-4.

Pérez & Arenas, et al. 2006

Pérez, J.; Arenas, M. & Gutierrez, C. (2006). “Semantics and Complexity of SPARQL”. *5th International Semantic Web Conference*. Ed. by I. Cruz; S. Decker; D. Allemang; C. Preist & D. Schwabe. Springer. (Athens, GA, USA, 5th - 9th Nob. 2006), pp. 30–43. DOI: 10.1007/11926078.

Perraudat & Lima, et al. 2019

Perraudat, A.; Lima, A.; Dauzère-Pérès, S. & Vialletelle, P. (2019). “A decision support system for a critical time constraint tunnel”. *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. Ed. by A. Jain & F. Levitov. (Saratoga Springs, NY, USA, 6th - 9th May. 2019), pp. 1–5. DOI: 10.1109/ASMC.2019.8791812.

Pham & Kuestenmacher, et al. 2023

Pham, A.-D.; Kuestenmacher, A. & Ploeger, P. G. (2023). “TSEM: Temporally-Weighted Spatiotemporal Explainable Neural Network for Multivariate Time Series”. *Proceedings of the 2023 Future of Information and Communication Conference (FICC)*. Ed. by K. Arai. Vol. 1. Springer. (San Francisco, CA, USA, 2nd - 3rd Mar. 2023), pp. 183–204. DOI: 10.1007/978-3-031-28073-3_13.

Pham & Afify 2005

Pham, D. & Afify, A. (2005). “Machine-learning techniques and their applications in manufacturing”. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 219.5, pp. 395–412. DOI: 10.1243/095440505X32274.

Pirovano & Ciccullo, et al. 2020

Pirovano, G.; Ciccullo, F.; Pero, M. & Rossi, T. (2020). “Scheduling batches with time constraints in wafer fabrication”. *International Journal of Operational Research* 37.1, pp. 1–31. DOI: 10.1504/IJOR.2020.104222.

Rao & Frtunikj 2018

Rao, Q. & Frtunikj, J. (2018). “Deep learning for self-driving cars: Chances and challenges”. *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*. Ed. by R. Stolle; S. Scholz & M. Broy. (Gothenburg, Sweden, 28th May. 2018), pp. 35–38. DOI: 10.1145/3194085.3194087.

Robinson & Giglio 1999

Robinson, J. K. & Giglio, R. (1999). “Capacity planning for semiconductor wafer fabrication

with time constraints between operations". *Proceedings of the 31st Conference on Winter Simulation: Simulation—a bridge to the future*. Ed. by P. Farrington & H. Nembhard. Vol. 1. (Phoenix, AZ, USA, 5th - 8th Dec. 1999), pp. 880–887. DOI: 10.1145/324138.324545.

Russell & Norvig 2021

Russell, S. & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. 4th ed. New York, NY: Pearson. DOI: 10.1145/201977.201989.

Sadeghi & Dauzère-Pérès, et al. 2015

Sadeghi, R.; Dauzère-Pérès, S.; Yugma, C. & Lepelletier, G. (2015). "Production control in semiconductor manufacturing with time constraints". *2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. Ed. by C. Ordonio & J. Schaller. (Saratoga Springs, NY, USA, 3rd - 6th May. 2015), pp. 29–33. DOI: 10.1109/ASMC.2015.7164446.

Schelter & Biessmann, et al. 2015

Schelter, S.; Biessmann, F.; Januschowski, T.; Salinas, D.; Seufert, S. & Szarvas, G. (2015). "On challenges in machine learning model management". *IEEE Data Engineering Bulletin* 1, p. 1.

Schelthoff & Jacobi, et al. 2022

Schelthoff, K.; Jacobi, C.; Schlosser, E.; Plohmann, D.; Janus, M. & Furmans, K. (2022). "Feature Selection for Waiting Time Predictions in Semiconductor Wafer Fabs". *IEEE Transactions on Semiconductor Manufacturing* 35.3, pp. 546–555. DOI: 10.1109/TSM.2022.3182855.

Schmitt & Monostori, et al. 2012

Schmitt, R.; Monostori, L.; Glöckner, H. & Viharos, Z. J. (2012). "Design and assessment of quality control loops for stable business processes". *CIRP Annals* 61.1, pp. 439–444. DOI: 10.1016/j.cirp.2012.03.055.

Scholl & Domaschke 2000

Scholl, W. & Domaschke, J. (2000). "Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations". *IEEE Transactions on Semiconductor Manufacturing* 13.3, pp. 273–277. DOI: 10.1109/66.857935.

Schuh & Wiendahl 1997

Schuh, G. & Wiendahl, H.-P. (1997). *Komplexität und Agilität: Steckt die Produktion in der Sackgasse?* 1st ed. Berlin: Springer. DOI: 10.1007/978-3-642-60841-4.

Schulz & Jacobi, et al. 2022

Schulz, B.; Jacobi, C.; Gisbrecht, A.; Evangelos, A.; Chan, C. W. & Gan, B. P. (2022). "Graph Representation and Embedding for Semiconductor Manufacturing Fab States". *2022 Winter Simulation Conference (WSC)*. Ed. by T. Chen; M. Fu & L. McGinnis. IEEE.

(Singapore, Singapore, 11th - 14th Dec. 2022), pp. 3382–3393. DOI: 10.1109/WSC57314.2022.10015297.

Shamsfard & Barforoush 2004

Shamsfard, M. & Barforoush, A. A. (2004). “Learning ontologies from natural language texts”. *International Journal of Human-Computer Studies* 60.1, pp. 17–63. DOI: 10.1016/j.ijhcs.2003.08.001.

Silver & Hubert, et al. 2018

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T., et al. (2018). “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. *Science* 362.6419, pp. 1140–1144. DOI: 10.1126/science.aar6404.

Studer & Benjamins, et al. 1998

Studer, R.; Benjamins, V. R. & Fensel, D. (1998). “Knowledge engineering: Principles and methods”. *Data & Knowledge Engineering* 25.1-2, pp. 161–197. DOI: 10.1016/S0169-023X(97)00056-6.

Su 2003

Su, L.-H. (2003). “A hybrid two-stage flowshop with limited waiting time constraints”. *Computers & Industrial Engineering* 44.3, pp. 409–424. DOI: 10.1016/S0360-8352(02)00216-4.

Sun & Choung, et al. 2005

Sun, D.-S.; Choung, Y.-I.; Lee, Y.-J. & Jang, Y.-C. (2005). “Scheduling and control for time-constrained processes in semiconductor manufacturing”. *ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005*. Ed. by J. Doran. (Piscataway, NJ, USA, 13th - 15th Sep. 2005), pp. 295–298. DOI: 10.1109/ISSM.2005.1513361.

Sutton & Barto 2018

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press.

Terkaj & Pedrielli, et al. 2012

Terkaj, W.; Pedrielli, G. & Sacco, M. (2012). “Virtual factory data model”. *Proceedings of the Workshop on Ontology and Semantic Web for Manufacturing 2012*. Ed. by D. Anstasiou; L. Ramos; S. Krima & Y.-J. Chen. (Graz, Austria, 24th Jul. 2012), pp. 29–43.

Terkaj & Tolio, et al. 2015

Terkaj, W.; Tolio, T. & Urgo, M. (2015). “A virtual factory approach for in situ simulation to support production and maintenance planning”. *CIRP Annals* 64.1, pp. 451–454. DOI: 10.1016/j.cirp.2015.04.121.

Terkaj & Urgo 2014

Terkaj, W. & Urgo, M. (2014). “Ontology-based modeling of production systems for design

and performance evaluation". *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*. Ed. by C. E. Pereira; L. Gomes & A. Colombo. IEEE. (Porto Alegre, Brazil, 27th - 30th Jul. 2014), pp. 748–753. DOI: 10.1109/INDIN.2014.6945606.

Terkaj & Urgo 2015

Terkaj, W. & Urgo, M. (2015). "A virtual factory data model as a support tool for the simulation of manufacturing systems". *Procedia CIRP* 28, pp. 137–142. DOI: 10.1016/j.procir.2015.04.023.

Toyoshima & Hasegawa, et al. 2013

Toyoshima, N.; Hasegawa, T.; Wu, K. & Arima, S. (2013). "Proactive control of engineering operations and lot loadings of product-mix and re-entrant in Q-time constraints processes". *2013 e-Manufacturing & Design Collaboration Symposium (eMDC)*. Ed. by L. Tuung. (Hsinchu, Taiwan, 6th Sep. 2013), pp. 1–4. DOI: 10.1109/eMDC.2013.6756046.

Tu & Chen & Liu 2010

Tu, Y. .-.; Chen, H. .-. & Liu, T. .-. (2010). "Shop-floor control for batch operations with time constraints in wafer fabrication". *International Journal of Industrial Engineering : Theory Applications and Practice* 17.2, pp. 142–155.

Tu & Chen 2011

Tu, Y.-M. & Chen, C.-L. (2011). "Model to determine the capacity of wafer fabrications for batch-serial processes with time constraints". *International Journal of Production Research* 49.10, pp. 2907–2923. DOI: 10.1080/00207541003730854.

Tu & Chen 2009a

Tu, Y.-M. & Chen, H.-N. (2009a). "Capacity planning with sequential two-level time constraints in the back-end process of wafer fabrication". *International Journal of Production Research* 47.24, pp. 6967–6979. DOI: 10.1080/00207540802415568.

Tu & Chen 2009b

Tu, Y.-M. & Chen, H.-N. (2009b). "Tool portfolio planning in the back-end process of wafer fabrication with sequential time constraints". *Journal of the Chinese Institute of Industrial Engineers* 26.1, pp. 60–69. DOI: 10.1080/10170660909509122.

Tu & Chen 2010

Tu, Y.-M. & Chen, H.-N. (2010). "Capacity planning with sequential time constraints under various control policies in the back-end of wafer fabrications". *Journal of the Operational Research Society* 61.8, pp. 1258–1264. DOI: 10.1057/jors.2009.36.

Tu & Liou 2006

Tu, Y.-M. & Liou, C.-S. (2006). "Capacity determination model with time constraints and batch processing in semiconductor wafer fabrication". *Journal of the Chinese Institute of Industrial Engineers* 23.3, pp. 192–199. DOI: 10.1080/10170660609509008.

Uçar & Le Dain, et al. 2020
Uçar, E.; Le Dain, M.-A. & Joly, I. (2020). “Digital technologies in circular economy transition: evidence from case studies”. *Procedia CIRP* 90, pp. 133–136. DOI: 10.1016/j.procir.2020.01.058.

Ueda & Fujii, et al. 2002
Ueda, K.; Fujii, N.; Hatono, I. & Kobayashi, M. (2002). “Facility layout planning using self-organization method”. *CIRP Annals* 51.1, pp. 399–402. DOI: 10.1016/S0007-8506(07)61546-7.

Uhlemann & Lehmann, et al. 2017
Uhlemann, T. H.-J.; Lehmann, C. & Steinhilper, R. (2017). “The digital twin: Realizing the cyber-physical production system for industry 4.0”. *Procedia CIRP* 61, pp. 335–340. DOI: 10.1016/j.procir.2016.11.152.

Valet & Altenmüller, et al. 2022
Valet, A.; Altenmüller, T.; Waschneck, B.; May, M. C.; Kuhnle, A. & Lanza, G. (2022). “Opportunistic maintenance scheduling with deep reinforcement learning”. *Journal of Manufacturing Systems* 64, pp. 518–534. DOI: 10.1016/j.jmsy.2022.07.016.

Vila 1994
Vila, L. (1994). “A survey on temporal reasoning in artificial intelligence”. *AI Communications* 7.1, pp. 4–28. DOI: 10.3233/AIC-1994-7102.

Wang & Liu 2013
Wang, C. & Liu, X.-B. (2013). “Integrated production planning and control: A multi-objective optimization model”. *Journal of Industrial Engineering and Management (JIEM)* 6.4, pp. 815–830. DOI: 10.3926/jiem.771.

Wang & Ju 2021
Wang, F. & Ju, F. (2021). “Decomposition-based real-time control of multi-stage transfer lines with residence time constraints”. *IISE Transactions* 53.9, pp. 943–959. DOI: 10.1080/24725854.2020.1803513.

Wang & Chien, et al. 2014
Wang, H.-K.; Chien, C.-F. & Gen, M. (2014). “Hybrid estimation of distribution algorithm with multiple subpopulations for semiconductor manufacturing scheduling problem with limited waiting-time constraint”. *2014 IEEE International Conference on Automation Science and Engineering (CASE)*. Ed. by X. Xie & M. C. Zhou. (Taipei, Taiwan, 18th - 22nd Aug. 2014), pp. 101–106. DOI: 10.1109/CoASE.2014.6899311.

Wang & Chien, et al. 2015
Wang, H.-K.; Chien, C.-F. & Gen, M. (2015). “An Algorithm of Multi-Subpopulation Parameters With Hybrid Estimation of Distribution for Semiconductor Scheduling With Constrained

Waiting Time". *IEEE Transactions on Semiconductor Manufacturing* 28.3, pp. 353–366. DOI: 10.1109/TSM.2015.2439054.

Wang & Hu, et al. 2020

Wang, J.; Hu, H.; Pan, C.; Zhou, Y. & Li, L. (2020). "Scheduling dual-arm cluster tools with multiple wafer types and residency time constraints". *IEEE/CAA Journal of Automatica Sinica* 7.3, pp. 776–789. DOI: 10.1109/JAS.2020.1003150.

Wang & Srivathsan, et al. 2018

Wang, M.; Srivathsan, S.; Huang, E. & Wu, K. (2018). "Job Dispatch Control for Production Lines With Overlapped Time Window Constraints". *IEEE Transactions on Semiconductor Manufacturing* 31.2, pp. 206–214. DOI: 10.1109/TSM.2018.2826530.

Wang & Usher 2004

Wang, Y. C. & Usher, J. M. (2004). "Learning policies for single machine job dispatching". *Robotics and Computer-Integrated Manufacturing* 20.6 SPEC. ISS. pp. 553–562. DOI: 10.1016/j.rcim.2004.07.003.

Waschneck & Altenmüller, et al. 2016

Waschneck, B.; Altenmüller, T.; Bauernhansl, T. & Kyek, A. (2016). "Production Scheduling in Complex Job Shops from an Industry 4.0 Perspective: A Review and Challenges in the Semiconductor Industry". *Proceedings of the SamI40 workshop at i-KNOW '16*. Ed. by R. Kern & R. Kaiser. (Graz, Austria, 18th - 19th Oct. 2016), pp. 1–12.

Wegener & Weikert, et al. 2021

Wegener, K.; Weikert, S.; Mayr, J.; Maier, M.; Ali Akbari, V. O. & Postel, M. (2021). "Operator integrated–concept for manufacturing intelligence". *Journal of Machine Engineering* 21, pp. 5–28. DOI: 10.36897/jme/144359.

Wiendahl & ElMaraghy, et al. 2007

Wiendahl, H.-P.; ElMaraghy, H. A.; Nyhuis, P.; Zäh, M. F.; Wiendahl, H.-H.; Duffie, N. & Brieke, M. (2007). "Changeable manufacturing-classification, design and operation". *CIRP Annals* 56.2, pp. 783–809. DOI: 10.1016/j.cirp.2007.10.003.

Wiendahl & Scholtissek 1994

Wiendahl, H.-P. & Scholtissek, P. (1994). "Management and control of complexity in manufacturing". *CIRP Annals* 43.2, pp. 533–540. DOI: 10.1016/S0007-8506(07)60499-5.

Wiendahl 1997

Wiendahl, H.-P. (1997). *Fertigungsregelung: Logistische Beherrschung von Fertigungsabläufen auf Basis des Trichtermodells*. München: Hanser.

Wiendahl & Breithaupt 1999

Wiendahl, H.-P. & Breithaupt, J.-W. (1999). "Modelling and controlling the dynamics of

production systems". *Production planning & control* 10.4, pp. 389–401. DOI: 10.1080/095372899233136.

Wolf & Debut, et al. 2020

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M., et al. (2020). "Transformers: State-of-the-art natural language processing". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu & D. Schlangen. (Online, 16th-20th Nov. 2020), pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wolfswinkel & Furtmueller, et al. 2013

Wolfswinkel, J. F.; Furtmueller, E. & Wilderom, C. P. (2013). "Using grounded theory as a method for rigorously reviewing literature". *European Journal of Information Systems* 22.1, pp. 45–55. DOI: 10.1057/ejis.2011.51.

Wu & Cheng, et al. 2012

Wu, C.-H.; Cheng, Y.-C.; Tang, P.-J. & Yu, J.-Y. (2012). "Optimal batch process admission control in tandem queueing systems with queue time constraint considerations". *Proceedings of the 2012 Winter Simulation Conference (WSC)*. Ed. by O. Rose & A. Uhrmacher. (Berlin, Germany, 9th - 12th Dec. 2012), pp. 1–6. DOI: 10.1109/WSC.2012.6465293.

Wu & Chien, et al. 2016

Wu, C.-H.; Chien, W.-C.; Chuang, Y.-T. & Cheng, Y.-C. (2016). "Multiple product admission control in semiconductor manufacturing systems with process queue time (PQT) constraints". *Computers & Industrial Engineering* 99, pp. 347–363. DOI: 10.1016/j.cie.2016.04.003.

Wu & Lin, et al. 2010

Wu, C.-H.; Lin, J. T. & Chien, W.-C. (2010). "Dynamic production control in a serial line with process queue time constraint". *International Journal of Production Research* 48.13, pp. 3823–3843. DOI: 10.1080/00207540902922836.

Wu & Lin, et al. 2012

Wu, C.-H.; Lin, J. T. & Chien, W.-C. (2012). "Dynamic production control in parallel processing systems under process queue time constraints". *Computers & Industrial Engineering* 63.1, pp. 192–203. DOI: 10.1016/j.cie.2012.02.003.

Wu & Zhao, et al. 2016

Wu, K.; Zhao, N.; Gao, L. & Lee, C. (2016). "Production control policy for tandem workstations with constant service times and queue time constraints". *International Journal of Production Research* 54.21, pp. 6302–6316. DOI: 10.1080/00207543.2015.1129468.

Wuest & Weimer, et al. 2016

Wuest, T.; Weimer, D.; Irgens, C. & Thoben, K.-D. (2016). "Machine learning in manufacturing: advantages, challenges, and applications". *Production & Manufacturing Research* 4.1, pp. 23–45. DOI: 10.1080/21693277.2016.1192517.

Wurster & Michel, et al. 2022
Wurster, M.; Michel, M.; May, M. C.; Kuhnle, A.; Stricker, N. & Lanza, G. (2022). “Modelling and condition-based control of a flexible and hybrid disassembly system with manual and autonomous workstations using reinforcement learning”. *Journal of Intelligent Manufacturing* 33.2, pp. 575–591. DOI: 10.1007/s10845-021-01863-3.

Xiao 2012
Xiao, H. (2012). *Introduction to Semiconductor Manufacturing Technology*. 2nd. Bellingham, WA, USA: SPIE Press.

Yadav & Yadav, et al. 2015
Yadav, N.; Yadav, A. & Kumar, M. (2015). “History of Neural Networks”. *An Introduction to Neural Network Methods for Differential Equations*. Dordrecht: Springer Netherlands, pp. 13–15. DOI: 10.1007/978-94-017-9816-7_2.

Yamak & Yujian, et al. 2019
Yamak, P. T.; Yujian, L. & Gadosey, P. K. (2019). “A comparison between ARIMA, LSTM, and GRU for time series forecasting”. *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. Ed. by S. Kundu & X. Xiao. (Sanya, China, 20th - 22nd Feb. 2019), pp. 49–55. DOI: 10.1145/3377713.3377722.

Yang & Ke, et al. 2015
Yang, K.-T.; Ke, L. & Shen, T. (2015). “Modeling and dispatching refinement for implantation to reduce the probability of tuning beam”. *26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. Ed. by C. Ordonio & J. Schaller. (Saratoga Springs, NY, USA, 3rd - 6th May. 2015), pp. 190–194. DOI: 10.1109/ASMC.2015.7164467.

Yang & Yu, et al. 2020
Yang, S.; Yu, X. & Zhou, Y. (2020). “LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example”. *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*. Ed. by G. Xu & M. A. Al-Khasawneh. (Guangzhou, China, 12th - 14th Jun. 2020), pp. 98–101. DOI: 10.1109/IWECAI50956.2020.00027.

Yu & Kim & Jung, et al. 2013
Yu, T.-S.; Kim, H.-J.; Jung, C. & Lee, T.-E. (2013). “Two-stage lot scheduling with waiting time constraints and due dates”. *Proceedings of the 2013 Winter Simulations Conference (WSC)*. Ed. by R. Hill & M. Kuhl. (Washington D.C., USA, 8th - 13th Dec. 2013), pp. 3630–3641. DOI: 10.1109/WSC.2013.6721724.

Yu & Kim & Lee 2017
Yu, T.-S.; Kim, H.-J. & Lee, T.-E. (2017). “Minimization of Waiting Time Variation in a Generalized Two-Machine Flowshop With Waiting Time Constraints and Skipping Jobs”.

IEEE Transactions on Semiconductor Manufacturing 30.2, pp. 155–165. DOI: 10.1109/TSM.2017.2662231.

Yugma & Dauzère-Pérès, et al. 2012

Yugma, C.; Dauzère-Pérès, S.; Artigues, C.; Derreumaux, A. & Sibille, O. (2012). "A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing". *International Journal of Production Research* 50.8, pp. 2118–2132. DOI: 10.1080/00207543.2011.575090.

Yurtsever & Kutanoglu, et al. 2009

Yurtsever, T.; Kutanoglu, E. & Johns, J. (2009). "Heuristic based scheduling system for diffusion in semiconductor manufacturing". *Proceedings of the 2009 Winter Simulation Conference (WSC)*. Ed. by A. Dunkin & R. Ingalls. (Austin, TX, USA, 13th - 16th Dec. 2009), pp. 1677–1685. DOI: 10.1109/WSC.2009.5429171.

Zhang & Pappert, et al. 2016

Zhang, T.; Pappert, F. S. & Rose, O. (2016). "Time bound control in a stochastic dynamic wafer fab". *Proceedings of the 2016 Winter Simulation Conference (WSC)*. Ed. by P. Frazier; T. Roeder & E. Zhou. (Washington D.C., USA, 11th - 14th Dec. 2016), pp. 2903–2911. DOI: 10.1109/WSC.2016.7822325.

Zhou & Lin, et al. 2019

Zhou, L.; Lin, C.; Hu, B. & Cao, Z. (2019). "A Cuckoo Search-Based Scheduling Algorithm for a Semiconductor Production Line with Constrained Waiting Time". *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. Ed. by W. Shen; J. Li & A. Matta. (Vancouver, Canada, 22nd - 26th Aug. 2019), pp. 338–343. DOI: 10.1109/COASE.2019.8842869.

Zhou & Pan, et al. 2017

Zhou, L.; Pan, S.; Wang, J. & Vasilakos, A. V. (2017). "Machine learning on big data: Opportunities and challenges". *Neurocomputing* 237, pp. 350–361. DOI: 10.1016/j.neucom.2017.01.026.

Zhou & Wu 2017

Zhou, Y. & Wu, K. (2017). "Heuristic simulated annealing approach for diffusion scheduling in a semiconductor Fab". *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. Ed. by R. Lee; X. Cui & S. Xu. (Wuhan, China, 24th - 26th May. 2017), pp. 785–789. DOI: 10.1109/ICIS.2017.7960099.

Zhu & Laptev 2017

Zhu, L. & Laptev, N. (2017). "Deep and confident prediction for time series at uber". *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. Ed. by G. Karypis & L. Miele. IEEE. (New Orleans, LO, USA, 18th -21st Nov. 2017), pp. 103–110. DOI: 10.1109/ICDMW.2017.19.

Ziarnetzky & Mönch, et al. 2017

Ziarnetzky, T.; Mönch, L.; Ponsignon, T. & Ehm, H. (2017). “Rolling horizon planning with engineering activities in semiconductor supply chains”. *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*. Ed. by X. Guan & Q. Zhao. IEEE. (Xi’an, China, 20th - 23rd Aug. 2017), pp. 1024–1025. DOI: 10.1109/COASE.2017.8256237.

List of Figures

2.1 Structure of the Fundamentals chapter with relations between sections. 6
2.2 Semiconductor manufacturing overview. Adapted from Mönch; Fowler; Dauzère-Pérès, et al. (2011) and Pe (2018). 7
2.3 Transistor and comparison of n- and p-semiconductors based on Pe (2018). . . 8
2.4 IC manufacturing steps. Adapted from Pe (2018) and Mönch; Fowler; Dauzère-Pérès, et al. (2011). 9
2.5 Frequent types of contamination in semiconductor manufacturing. 13
2.6 Wafer fabrication in a semiconductor fab. Adapted from Xiao (2012). 15
2.7 Wafer flowchart for a wafer fab. Adapted from Mönch; Fowler; Dauzère-Pérès, et al. (2011). 16
2.8 Machine dedications and machine failures influence cycle times and the processing order. 17
2.9 Time-constraint types limiting the transition time between two operations classified based on Klemmt & Mönch (2012) and Wang; Srivathsan, et al. (2018). . . 18
2.10 Complexities in wafer fabrication based on Mönch; Fowler & Mason (2013), Xiao (2012) and Valet et al. (2022). 20
2.11 Production Planning and Control scope in a market economy, based on Wiendahl; ElMaraghy, et al. (2007) 23
2.12 Levels of Production Planning and Control (PPC), based on Schuh & Wiendahl (1997); Mönch; Fowler & Mason (2013) and Nickel et al. (2014) 24
2.13 Traditional complexity drivers for PPC in manufacturing, based on Wiendahl & Scholtissek (1994) 25
2.14 Characteristics of complex job shops based on Waschneck et al. (2016) and Mönch; Fowler & Mason (2013) 27
2.15 Development of complex jobs and complexity comparison with preceding production approaches. 28
2.16 Semiconductor PPC hierarchy according to Mönch; Fowler & Mason (2013). . . 29
2.17 Complex job shops strongest PPC complexity drivers. 31
2.18 Illustrative comparison of exact and approximate quantitative optimization methods (Machine Learning and Heuristics) with respect to suitable size of problems, quality of solution and required computational effort based on Nickel et al. (2014). 32
2.19 Morphological box of a selection of optimization problem types based on Nickel et al. (2014). 33
2.20 Artificial Intelligence and relevant subsets based on Russell & Norvig (2021). . 36

2.21 Examplary knowledge graph as a direct edge-labeled graph for lot data in a CMP process. . . . 38
2.22 Exemplary visualization of the architecture of a fully connected neural network. 42
2.23 Recurrent neural network architecture and unfolding over time based on (Goodfellow et al. 2016). . . . 44
2.24 Comparison of the cell structure of an LSTM and GRU model. . . . 44
2.25 Bulls-eye visualization of the trade-off in bias and variance based on (Gudivada et al. 2017). . . . 47
2.26 Benefit of using prediction intervals over pure point estimation based on May; Maucher, et al. (2021). . . . 51
2.27 Interconnection of an ontology and knowledge graph based digital twin with its physical counterpart based on May; Kiefer; Kuhnle & Lanza (2022). . . . 56
2.28 Evolution and increased benefit from simulation to foresighted and ultimately ontology-based Digital Twin for a production system. . . . 58

4.1 Methodological approach to obtain the intelligent production for time-constraint gate keeping in complex job shops. . . . 76
4.2 Organization of Section 4.2. . . . 79
4.3 Categorization of system elements in the simulation. . . . 82
4.4 Concept of a simplified product modeled in the simulation based on May; Kiefer; Kuhnle & Lanza (2022). . . . 83
4.5 Concept of a simplified processing resource modeled in the simulation based on May; Kiefer; Kuhnle & Lanza (2022). . . . 84
4.6 Concept of events modeled in the simulation based on May; Kiefer; Kuhnle & Lanza (2022). . . . 85
4.7 Events during an exemplary simulation run based on May; Kiefer; Kuhnle & Lanza (2022). . . . 87
4.8 Product process flow chart with the regarded PPC decisions and the focus on time-constraint gate keeping decisions. . . . 88
4.9 Graph based transitional modeling approach as a complex job shop abstraction with an exemplary transition based on May; Maucher, et al. (2021). . . . 92
4.10 Sparse transitional model graph and the temporal graph including material flow represented through boxes on the transition based on May; Maucher, et al. (2021). 93
4.11 Organization of Section 4.3. . . . 95
4.12 Product process flow chart with the regarded PPC decisions and the focus on time-constraint gate keeping decisions. . . . 96
4.13 Decision framework for time-constrained lots gate keeping decision. . . . 97
4.14 General rollout procedure. . . . 99

4.15 Parallelization of rollout behavior during a foresight period. 101
4.16 Exemplary auto-correlation between consecutive transition times on the same transition based on May; Maucher, et al. (2021). 102
4.17 Transitional model based on May; Maucher, et al. (2021) and May; Behnen, et al. (2021). 103
4.18 Organization of Section 4.4. 105
4.19 General decision process for gate keeping with time-constraint adherence prediction. 106
4.20 Using the foresighted digital twin to predict time-constraint adherence for the gate keeping decision. 107
4.21 Transitional model based gate keeping based on one-sided prediction interval. 110
4.22 Three approaches to obtain basic prediction interval estimators (a) Chebyshev, (b) assumed Normal distribution or (c) assumed Lognormal distribution. 116
4.23 Observed transition time distribution and logarithmic transformation. 117
4.24 Kolmogorov-Smirnov test comparison for large sample size and selection of the first 30 records based on May; Maucher, et al. (2021). 118
4.25 Comparison of a neural network with droupout applied during this training step and a general fully conected neural network. 123
4.26 Organization of Section 4.5. 124
4.27 Single random sampling rollout point estimator from the foresighted digital twin for an exemplary transition with a sequence of 45 lot transitions. 125
4.28 Exemplary ARIMA point estimator for a transition. 127
4.29 Development of the loss for the training and validation datasets over the training epochs. 128
4.30 Exemplary neural network point estimator for a transition trained on the proposed approach. 129
4.31 Comparing expected and observed prediction interval coverages based on May; Maucher, et al. (2021) and May; Behnen, et al. (2021). 130
4.32 Residual analysis and prediction ability of a feed-forward neural network based on May; Behnen, et al. (2021). 133
4.33 Summary of own approach and the implemented process elements. 133

5.1 Exemplary observed transition time and split into training, validation and test time frames. 138
5.2 Exemplary observed transition time, predicted transition time, time limits and the prediction interval at the designated coverage, where actual violations can be seen whenever the observed transition time exceeds the time limit. 142

5.3 Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with an ARIMA model. 146
5.4 Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a NN model. 147
5.5 Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a LSTM model. 147
5.6 Exemplary observed transition with point estimators and prediction intervals at the maximum coverage of 80% with a GRU model. 148

A1.1 Comparison of transition time distribution in a boxplot for transitions (left) and operations (right) from the real-world example. X
A1.2 Comparison of transition times for two equipment under the influence of breakdowns during the lots being in transit (left) and the influence of a current breakdown on lots to be dispatched. XI
A2.1 Visualization of an exemplary transition and the ARIMA point estimator for one randomly selected week. XII
A2.2 Visualization of an exemplary transition and the LSTM point estimator for one randomly selected week. XIII
A3.1 Exemplary transition schowing adherence prediction with final ARIMA model at given coverage of 80% based on May; Maucher, et al. (2021) XIV

List of Tables

2.1 Most relevant KPIs as performance measures in production planning and control 21

3.1 Overview of relevant research on capacity planning and intelligent production control approaches for time-constrained complex job shops. 73
3.2 Overview of relevant research on scheduling for time-constrained complex job shops. 74
3.3 Overview of relevant research on dispatching for time-constrained complex job shops. 75

4.1 Comparing coverage levels and required sample sizes for the logarithmic model.131
4.2 Custom one-sided Winkler Loss for the regarded period and different distributional assumptions. 132

5.1 Recall, precision and accuracy calculation. 140
5.2 Confusion matrix for the binary classification of time-constraint violation and adherence of the foresighted digital twin model in the reduced test data set. . . 143
5.3 Recall, precision and accuracy for the foresighted digital twin based approach. 143
5.4 Prediction interval coverage probability and mean prediction interval width of the NN, LSTM and GRU models. 146
5.5 Recall, precision and accuracy comparison for the selected prediction models in the transitional modeling approach with the preferred 80% coverage interval. 148

Appendices

A1 Additional data analysis of the real-world semiconductor manufacturing dataset

Within a complex job shop the high degree of stochastic influences and convoluted material flow leads to high variations in the actual transition time. The transition time denotes the time a lot spends between two operations and which may or may not be limited by time-constraints. Regarding the real-world semiconductor manufacturing plant which is regarded for the validation, the actual transition times observed can vary significantly as the depend on the processing times, distances and criticality of certain equipment and operations. Figure A1.1 visualizes the large variation in transition times, which underlines the need to intelligently apply production control to enforce time-constraint adherence. In a similar vein, the operations that are required subsequently to arriving at the transitions destination from the identical sending equipment heavily influences the observed transition times as shown in the right part of Figure A1.1. By regarding transitions individually and by leveraging time series model the proposed approach aims to develop and intelligent production control for these time-constrained complex job shops.

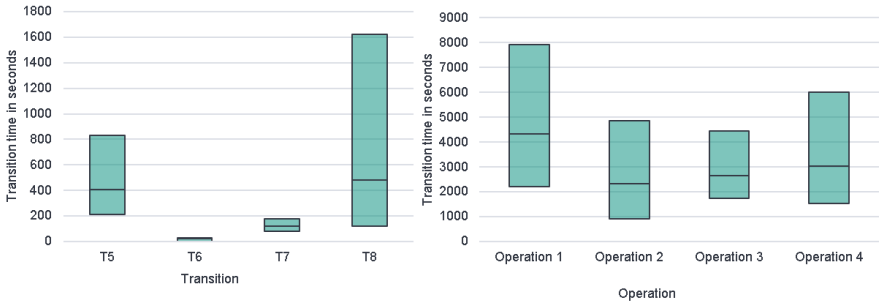

Figure A1.1: Comparison of transition time distribution in a boxplot for transitions (left) and operations (right) from the real-world example.

Besides the geospatial and operational influence, the stochastic behavior of the underlying equipment used is exhibiting a large influence on the actual transition time observed and, hence, on the number of time-constraint violations. Figure A1.2 visualizes the influence of a breakdown of the target equipment on the transition time of a corresponding transition. In case a breakdown occurs after the lot has been started operating or while it is in transit, the transition time is heavily influenced as shown on the left. Therefore, using predictive

approaches such as the foresighted digital twin can leverage good expectations about such events. On the other hand, there is a much lighter influence on the transition time if a current breakdown is known before dispatching the respective lot.

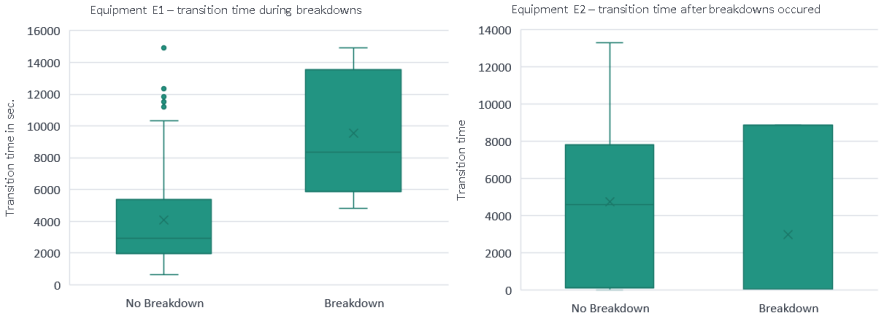

Figure A1.2: Comparison of transition times for two equipment under the influence of breakdowns during the lots being in transit (left) and the influence of a current breakdown on lots to be dispatched.

A2 Additional experimental evaluation of transition time prediction

Accurate predictions of the transition time are key to well adjusted prediction intervals to intelligently control the gate keeping production control decision for improving time-constraint adherence in complex job shops. Regarding one individual transition (T9) as in the following Figure A2.1 illustrates the high degree of variability in transition times, just for one transition. Using the time series prediction models a predictor is built to predict these timings as visualized. Clearly, the ARIMA model is not fully capable of perfectly predicting the transition time, thus, magnifying the need to use the prediction interval extension to not only regard one value but the cumulative probability of realizations at or below the time-constraint time limit. The density of transition usage changes over time visualizing the high degree of change present in a complex job shop and in this particular real-world semiconductor manufacturing plant.

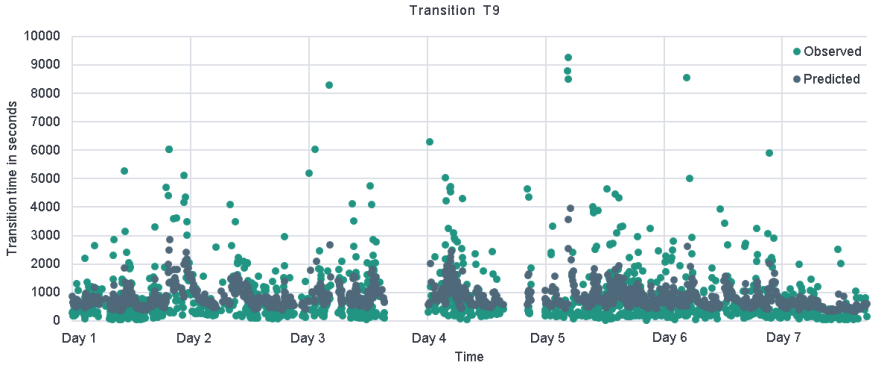

Figure A2.1: Visualization of an exemplary transition and the ARIMA point estimator for one randomly selected week.

In a similar vein Figure A2.2 shows the same transition during the same week and highlights the LSTM made predictions. From visual comparison the LSTM seems more capable of predicting the exact transition times. However, for the prediction interval the overall interval quality and not solely the point estimator is decisive. Therefore, as discussed in Chapter 5 the overall model needs to be regarded, where the prediction interval is additionally influenced

by the width that is depending on the quality and number of parameters in the prediction model.

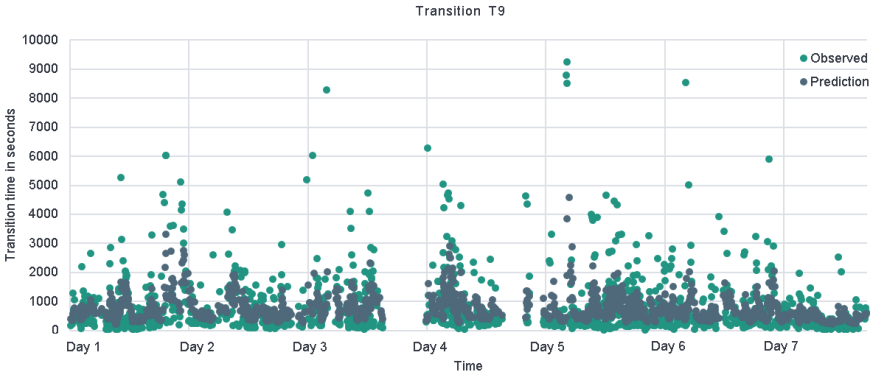

Figure A2.2: Visualization of an exemplary transition and the LSTM point estimator for one randomly selected week.

A3 Additional experimental evaluation of time-constraint adherence prediction

As an alternative to the pre-selected coverage of 80% an ex post evaluation of the individual coverage levels is possible. Figure A3.1 shows an exemplary transition where all time-constrained lots have been regarded with their individual adherence probability sorted in a descending order. Clearly, based on the 80% coverage interval all three time-constraint violations would be preventable. A lower coverage of 75% would lead to one violation. This ex post evaluation, however, is hardly suitable for selecting the coverage as an individual transition specific coverage would have to be identified, stored and continuously updated. For future work this is promising.

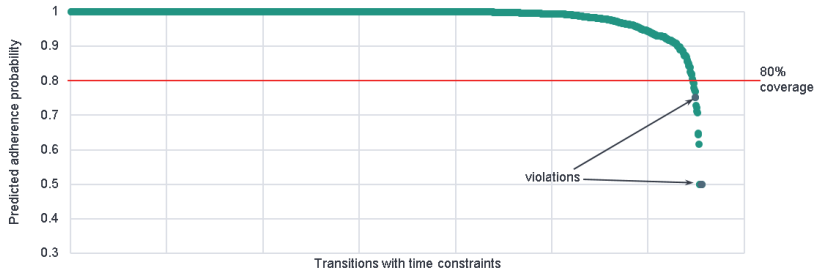

Figure A3.1: Exemplary transition schowing adherence prediction with final ARIMA model at given coverage of 80% based on May; Maucher, et al. (2021)

Forschungsberichte aus dem wbk
Institut für Produktionstechnik
Karlsruher Institut für Technologie (KIT)

Bisher erschienene Bände:

Band 0
Dr.-Ing. Wu Hong-qi

Adaptive Volumenstromregelung mit Hilfe von drehzahlgeregelten Elektroantrieben

Band 1
Dr.-Ing. Heinrich Weiß

Fräsen mit Schneidkeramik - Verhalten des System Werkzeugmaschine-Werkzeug-Werkstück und Prozessanalyse

Band 2
Dr.-Ing. Hans-Jürgen Stierle

Entwicklung und Untersuchung hydrostatischer Lager für die Axialkolbenmaschine

Band 3
Dr.-Ing. Herbert Hörner

Untersuchung des Geräuschverhaltens druckgeregelter Axialkolbenpumpen

Band 4
Dr.-Ing. Rolf-Dieter Brückbauer

Digitale Drehzahlregelung unter der besonderen Berücksichtigung von Quantisierungseffekten

Band 5
Dr.-Ing. Gerhard Staiger

Graphisch interaktive NC-Programmierung von Drehteilen im Werkstattbereich

Band 6
Dr.-Ing. Karl Peters

Ein Beitrag zur Berechnung und Kompensation von Positionierfehlern an Industrierobotern

Band 7
Dr.-Ing. Paul Stauss

Automatisierte Inbetriebnahme und Sicherung der Zuverlässigkeit und Verfügbarkeit numerisch gesteuerter Fertigungseinrichtungen

Band 8
Dr.-Ing. Günter Möckesch

Konzeption und Realisierung eines strategischen, integrierten Gesamtplanungs- und -bearbeitungssystems zur Optimierung der Drehteilorganisation für auftragsbezogene Drehereien

Band 9
Dr.-Ing. Thomas Oestreicher

Rechnergestützte Projektierung von Steuerungen

Band 10
Dr.-Ing. Thomas Selinger

Teilautomatisierte werkstattnahe NC-Programmerstellung im Umfeld einer integrierten Informationsverarbeitung

Band 11
Dr.-Ing. Thomas Buchholz

Prozessmodell Fräsen, Rechnerunterstützte Analyse, Optimierung und Überwachung

Band 12
Dr.-Ing. Bernhard Reichling

Lasergestützte Positions- und Bahnvermessung von Industrierobotern

Band 13
Dr.-Ing. Hans-Jürgen Lesser

Rechnergestützte Methoden zur Auswahl anforderungsgerechter Verbindungselemente

Band 14
Dr.-Ing. Hans-Jürgen Lauffer

Einsatz von Prozessmodellen zur rechnerunterstützten Auslegung von Räumwerkzeugen

Band 15
Dr.-Ing. Michael C. Wilhelm

Rechnergestützte Prüfplanung im Informationsverbund moderner Produktionssysteme

Band 16
Dr.-Ing. Martin Ochs

Entwurf eines Programmsystems zur wissensbasierten Planung und Konfigurierung

Band 17
Dr.-Ing. Heinz-Joachim Schneider

Erhöhung der Verfügbarkeit von hochautomatisierten Produktionseinrichtungen mit Hilfe der Fertigungsleittechnik

Band 18
Dr.-Ing. Hans-Reiner Ludwig

Beanspruchungsanalyse der Werkzeugschneiden beim Stirnplanfräsen

Band 19
Dr.-Ing. Rudolf Wieser

Methoden zur rechnergestützten Konfigurierung von Fertigungsanlagen

Band 20
Dr.-Ing. Edgar Schmitt

Werkstattsteuerung bei wechselnder Auftragsstruktur

Band 21
Dr.-Ing. Wilhelm Enderle

Verfügbarkeitssteigerung automatisierter Montagesysteme durch selbsttätige Behebung prozessbedingter Störungen

Band 22
Dr.-Ing. Dieter Buchberger

Rechnergestützte Strukturplanung von Produktionssystemen

Band 23
Prof. Dr.-Ing. Jürgen Fleischer

Rechnerunterstützte Technologieplanung für die flexibel automatisierte Fertigung von Abkantteilen

Band 24
Dr.-Ing. Lukas Loeffler

Adaptierbare und adaptive Benutzerschnittstellen

Band 25
Dr.-Ing. Thomas Friedmann

Integration von Produktentwicklung und Montageplanung durch neue rechnergestützte Verfahren

Band 26
Dr.-Ing. Robert Zurrin

Variables Formhonen durch rechnergestützte Hornprozesssteuerung

Band 27
Dr.-Ing. Karl-Heinz Bergen

Langhub-Innenrundhonen von Grauguss und Stahl mit einem elektromechanischem Vorschubsystem

Band 28
Dr.-Ing. Andreas Liebisch

Einflüsse des Festwalzens auf die Eigenspannungsverteilung und die Dauerfestigkeit einsatzgehärteter Zahnräder

Band 29
Dr.-Ing. Rolf Ziegler

Auslegung und Optimierung schneller Servopumpen

Band 30
Dr.-Ing. Rainer Bartl

Datenmodellgestützte Wissensverarbeitung zur Diagnose und Informationsunterstützung in technischen Systemen

Band 31
Dr.-Ing. Ulrich Golz

Analyse, Modellbildung und Optimierung des Betriebsverhaltens von Kugelgewindetrieben

Band 32
Dr.-Ing. Stephan Timmermann

Automatisierung der Feinbearbeitung in der Fertigung von Hohlformwerkzeugen

Band 33
Dr.-Ing. Thomas Noe

Rechnergestützter Wissenserwerb zur Erstellung von Überwachungs- und Diagnoseexpertensystemen für hydraulische Anlagen

Band 34
Dr.-Ing. Ralf Lenschow

Rechnerintegrierte Erstellung und Verifikation von Steuerungsprogrammen als Komponente einer durchgängigen Planungsmethodik

Band 35
Dr.-Ing. Matthias Kallabis

Räumen gehärteter Werkstoffe mit kristallinen Hartstoffen

Band 36
Dr.-Ing. Heiner-Michael Honeck

Rückführung von Fertigungsdaten zur Unterstützung einer fertigungsgerechten Konstruktion

Band 37
Dr.-Ing. Manfred Rohr

Automatisierte Technologieplanung am Beispiel der Komplettbearbeitung auf Dreh-/Fräszellen

Band 38
Dr.-Ing. Martin Steuer

Entwicklung von Softwarewerkzeugen zur wissensbasierten Inbetriebnahme von komplexen Serienmaschinen

Band 39
Dr.-Ing. Siegfried Beichter

Rechnergestützte technische Problemlösung bei der Angebotserstellung von flexiblen Drehzellen

Band 40
Dr.-Ing. Thomas Steitz

Methodik zur marktorientierten Entwicklung von Werkzeugmaschinen mit Integration von funktionsbasierter Strukturierung und Kostenschätzung

Band 41
Dr.-Ing. Michael Richter

Wissensbasierte Projektierung elektrohydraulischer Regelungen

Band 42
Dr.-Ing. Roman Kuhn

Technologieplanungssystem Fräsen. Wissensbasierte Auswahl von Werkzeugen, Schneidkörpern und Schnittbedingungen für das Fertigingsverfahren Fräsen

Band 43
Dr.-Ing. Hubert Klein

Rechnerunterstützte Qualitätssicherung bei der Produktion von Bauteilen mit frei geformten Oberflächen

Band 44
Dr.-Ing. Christian Hoffmann

Konzeption und Realisierung eines fertigungsintegrierten Koordinatenmessgerätes

Band 45
Dr.-Ing. Volker Frey

Planung der Leittechnik für flexible Fertigungsanlagen

Band 46
Dr.-Ing. Achim Feller

Kalkulation in der Angebotsphase mit dem selbsttätig abgeleiteten Erfahrungswissen der Arbeitsplanung

Band 47
Dr.-Ing. Markus Klaiber

Produktivitätssteigerung durch rechnerunterstütztes Einfahren von NC-Programmen

Band 48
Dr.-Ing. Roland Minges

Verbesserung der Genauigkeit beim fünfachsigen Fräsen von Freiformflächen

Band 49
Dr.-Ing. Wolfgang Bernhart

Beitrag zur Bewertung von Montagevarianten: Rechnergestützte Hilfsmittel zur kostenorientierten, parallelen Entwicklung von Produkt und Montagesystem

Band 50
Dr.-Ing. Peter Ganghoff

Wissensbasierte Unterstützung der Planung technischer Systeme: Konzeption eines Planungswerkzeuges und exemplarische Anwendung im Bereich der Montagesystemplanung

Band 51
Dr.-Ing. Frank Maier

Rechnergestützte Prozessregelung beim flexiblen Gesenkbiegen durch Rückführung von Qualitätsinformationen

Band 52
Dr.-Ing. Frank Debus

Ansatz eines rechnerunterstützten Planungsmanagements für die Planung in verteilten Strukturen

Band 53
Dr.-Ing. Joachim Weinbrecht

Ein Verfahren zur zielorientierten Reaktion auf Planabweichungen in der Werkstattregelung

Band 54
Dr.-Ing. Gerd Herrmann

Reduzierung des Entwicklungsaufwandes für anwendungsspezifische Zellenrechnersoftware durch Rechnerunterstützung

Band 55
Dr.-Ing. Robert Wassmer

Verschleissentwicklung im tribologischen System Fräsen: Beiträge zur Methodik der Prozessmodellierung auf der Basis tribologisher Untersuchungen beim Fräsen

Band 56
Dr.-Ing. Peter Uebelhoer

Inprocess-Geometriemessung beim Honen

Band 57
Dr.-Ing. Hans-Joachim Schelberg

Objektorientierte Projektierung von SPS-Software

Band 58
Dr.-Ing. Klaus Boes

Integration der Qualitätsentwicklung in featurebasierte CAD/CAM-Prozessketten

Band 59
Dr.-Ing. Martin Schreiber

Wirtschaftliche Investitionsbewertung komplexer Produktionssysteme unter Berücksichtigung von Unsicherheit

Band 60
Dr.-Ing. Ralf Steuernagel

Offenes adaptives Engineering-Werkzeug zur automatisierten Erstellung von entscheidungsunterstützenden Informationssystemen

Band 62
Dr.-Ing. Uwe Schauer

Qualitätsorientierte Feinbearbeitung mit Industrierobotern: Regelungsansatz für die Freiformflächenfertigung des Werkzeug- und Formenbaus

Band 63
Dr.-Ing. Simone Loeper

Kennzahlengestütztes Beratungssystem zur Verbesserung der Logistikleistung in der Werkstattfertigung

Band 64
Dr.-Ing. Achim Raab

Räumen mit hartstoffbeschichteten HSS-Werkzeugen

Band 65,
Dr.-Ing. Jan Erik Burghardt

Unterstützung der NC-Verfahrenskette durch ein bearbeitungs-elementorientiertes, lernfähiges Technologieplanungssystem

Band 66
Dr.-Ing. Christian Tritsch

Flexible Demontage technischer Gebrauchsgüter: Ansatz zur Planung und (teil-)automatisierten Durchführung industireller Demontageprozesse

Band 67
Dr.-Ing. Oliver Eitrich

Prozessorientiertes Kostenmodell für die entwicklungsbegleitende Vorkalkulation

Band 68
Dr.-Ing. Oliver Wilke

Optimierte Antriebskonzepte für Räummaschinen - Potentiale zur Leistungs-steigerung

Band 69
Dr.-Ing. Thilo Sieth

Rechnergestützte Modellierungsmethodik zerspantechnologischer Prozesse

Band 70
Dr.-Ing. Jan Linnenbuerger

Entwicklung neuer Verfahren zur automatisierten Erfassung der geometrischen Abweichungen an Linearachsen und Drehschwenkköpfen

Band 71
Dr.-Ing. Mathias Klimmek

Fraktionierung technischer Produkte mittels eines frei beweglichen Wasserstrahlwerkzeuges

Band 72
Dr.-Ing. Marko Hartel

Kennzahlenbasiertes Bewertungssystem zur Beurteilung der Demontage- und Recyclingeignung von Produkten

Band 73
Dr.-Ing. Jörg Schaupp

Wechselwirkung zwischen der Maschinen- und Hauptspindelantriebsdynamik und dem Zerspanprozess beim Fräsen

Band 74
Dr.-Ing. Bernhard Neisius

Konzeption und Realisierung eines experimentellen Telemanipulators für die Laparoskopie

Band 75
Dr.-Ing. Wolfgang Walter

Erfolgsversprechende Muster für betriebliche Ideenfindungsprozesse. Ein Beitrag zur Steigerung der Innovationsfähigkeit

Band 76
Dr.-Ing. Julian Weber

Ein Ansatz zur Bewertung von Entwicklungsergebnissen in virtuellen Szenarien

Band 77
Dr.-Ing. Dipl. Wirtsch.-Ing. Markus Posur

Unterstützung der Auftragsdurchsetzung in der Fertigung durch Kommunikation über mobile Rechner

Band 78
Dr.-Ing. Frank Fleissner

Prozessorientierte Prüfplanung auf Basis von Bearbeitungsobjekten für die Kleinserienfertigung am Beispiel der Bohr- und Fräsbearbeitung

Band 79
Dr.-Ing. Anton Haberkern

Leistungsfähigere Kugelgewindetriebe durch Beschichtung

Band 80
Dr.-Ing. Dominik Matt

Objektorientierte Prozess- und Strukturinnovation (OPUS)

Band 81
Dr.-Ing. Jürgen Andres

Robotersysteme für den Wohnungsbau: Beitrag zur Automatisierung des Mauerwerkabaus und der Elektroinstallation auf Baustellen

Band 82
Dr.-Ing. Dipl.Wirtschaftsing. Simone Riedmiller

Der Prozesskalender - Eine Methodik zur marktorientierten Entwicklung von Prozessen

Band 83
Dr.-Ing. Dietmar Tilch

Analyse der Geometrieparameter von Präzisionsgewinden auf der Basis einer Least-Squares-Estimation

Band 84
Dr.-Ing. Dipl.-Kfm. Oliver Stiefbold

Konzeption eines reaktionsschnellen Planungssystems für Logistikketten auf Basis von Software-Agenten

Band 85
Dr.-Ing. Ulrich Walter

Einfluss von Kühlschmierstoff auf den Zerspanprozess beim Fräsen: Beitrag zum Prozessverständniss auf Basis von zerspantechnischen Untersuchungen

Band 86
Dr.-Ing. Bernd Werner

Konzeption von teilautonomer Gruppenarbeit unter Berücksichtigung kultureller Einflüsse

Band 87
Dr.-Ing. Ulf Osmers

Projektieren Speicherprogrammierbarer Steuerungen mit Virtual Reality

Band 88
Dr.-Ing. Oliver Doerfel

Optimierung der Zerspantechnik beim Fertigungsverfahren Wälzstossen: Analyse des Potentials zur Trockenbearbeitung

Band 89
Dr.-Ing. Peter Baumgartner

Stufenmethode zur Schnittstellengestaltung in der internationalen Produktion

Band 90
Dr.-Ing. Dirk Vossmann

Wissensmanagement in der Produktentwicklung durch Qualitätsmethodenverbund und Qualitätsmethodenintegration

Band 91
Dr.-Ing. Martin Plass

Beitrag zur Optimierung des Honprozesses durch den Aufbau einer Honprozessregelung

Band 92
Dr.-Ing. Titus Konold

Optimierung der Fünfachsfräsbearbeitung durch eine kennzahlenunterstützte CAM-Umgebung

Band 93
Dr.-Ing. Jürgen Brath

Unterstützung der Produktionsplanung in der Halbleiterfertigung durch risikoberücksichtigende Betriebskennlinien

Band 94
Dr.-Ing. Dirk Geisinger

Ein Konzept zur marktorientierten Produktentwicklung

Band 95
Dr.-Ing. Marco Lanza

Entwurf der Systemunterstützung des verteilten Engineering mit Axiomatic Design

Band 96
Dr.-Ing. Volker Hüntrup

Untersuchungen zur Mikrostrukturierbarkeit von Stählen durch das Fertigungsverfahren Fräsen

Band 97
Dr.-Ing. Frank Reinboth

Interne Stützung zur Genauigkeitsverbesserung in der Inertialmesstechnik: Beitrag zur Senkung der Anforderungen an Inertialsensoren

Band 98
Dr.-Ing. Lutz Trender

Entwicklungsintegrierte Kalkulation von Produktlebenszykluskosten auf Basis der ressourcenorientierten Prozesskostenrechnung

Band 99
Dr.-Ing. Cornelia Kafka

Konzeption und Umsetzung eines Leitfadens zum industriellen Einsatz von Data-Mining

Band 100
Dr.-Ing. Gebhard Selinger

Rechnerunterstützung der informellen Kommunikation in verteilten Unternehmensstrukturen

Band 101
Dr.-Ing. Thomas Windmüller

Verbesserung bestehender Geschäftsprozesse durch eine mitarbeiterorientierte Informationsversorgung

Band 102
Dr.-Ing. Knud Lembke

Theoretische und experimentelle Untersuchung eines bistabilen elektrohydraulischen Linearantriebs

Band 103
Dr.-Ing. Ulrich Thies

Methode zur Unterstützung der variantengerechten Konstruktion von industriell eingesetzten Kleingeräten

Band 104
Dr.-Ing. Andreas Schmälzle

Bewertungssystem für die Generalüberholung von Montageanlagen –Ein Beitrag zur wirtschaftlichen Gestaltung geschlossener Facility- Managment-Systeme im Anlagenbau

Band 105
Dr.-Ing. Thorsten Frank

Vergleichende Untersuchungen schneller elektromechanischer Vorschubachsen mit Kugelgewindetrieb

Band 106
Dr.-Ing. Achim Agostini

Reihenfolgeplanung unter Berücksichtigung von Interaktionen: Beitrag zur ganzheitlichen Strukturierung und Verarbeitung von Interaktionen von Bearbeitungsobjekten

Band 107
Dr.-Ing. Thomas Barrho

Flexible, zeitfenstergesteuerte Auftragseinplanung in segmentierten Fertigungsstrukturen

Band 108
Dr.-Ing. Michael Scharer

Quality Gate-Ansatz mit integriertem Risikomanagement

Band 109
Dr.-Ing. Ulrich Suchy

Entwicklung und Untersuchung eines neuartigen Mischkopfes für das Wasser Abrasivstrahlschneiden

Band 110
Dr.-Ing. Sellal Mussa

Aktive Korrektur von Verlagerungsfehlern in Werkzeugmaschinen

Band 111
Dr.-Ing. Andreas Hühsam

Modellbildung und experimentelle Untersuchung des Wälzschälprozesses

Band 112
Dr.-Ing. Axel Plutowsky

Charakterisierung eines optischen Messsystems und den Bedingungen des Arbeitsraums einer Werkzeugmaschine

Band 113
Dr.-Ing. Robert Landwehr

Konsequent dezentralisierte Steuerung mit Industrial Ethernet und offenen Applikationsprotokollen

Band 114
Dr.-Ing. Christoph Dill

Turbulenzreaktionsprozesse

Band 115
Dr.-Ing. Michael Baumeister

Fabrikplanung im turbulenten Umfeld

Band 116
Dr.-Ing. Christoph Gönnheimer

Konzept zur Verbesserung der Elektromagnetischen Verträglichkeit (EMV) in Produktionssystemen durch intelligente Sensor/Aktor-Anbindung

Band 117
Dr.-Ing. Lutz Demuß

Ein Reifemodell für die Bewertung und Entwicklung von Dienstleistungsorganisationen: Das Service Management Maturity Modell (SMMM)

Band 118
Dr.-Ing. Jörg Söhner

Beitrag zur Simulation zerspanungstechnologischer Vorgänge mit Hilfe der Finite-Element-Methode

Band 119
Dr.-Ing. Judith Elsner

Informationsmanagement für mehrstufige Mikro-Fertigungsprozesse

Band 120
Dr.-Ing. Lijing Xie

Estimation Of Two-dimension Tool Wear Based On Finite Element Method

Band 121
Dr.-Ing. Ansgar Blessing

Geometrischer Entwurf mikromechatronischer Systeme

Band 122
Dr.-Ing. Rainer Ebner

Steigerung der Effizienz mehrachsiger Fräsprozesse durch neue Planungsmethoden mit hoher Benutzerunterstützung

Band 123
Dr.-Ing. Silja Klinkel

Multikriterielle Feinplanung in teilautonomen Produktionsbereichen – Ein Beitrag zur produkt- und prozessorientierten Planung und Steuerung

Band 124
Dr.-Ing. Wolfgang Neithardt

Methodik zur Simulation und Optimierung von Werkzeugmaschinen in der Konzept- und Entwurfsphase auf Basis der Mehrkörpersimulation

Band 125
Dr.-Ing. Andreas Mehr

Hartfeinbearbeitung von Verzahnungen mit kristallinen diamantbeschichteten Werkzeugen beim Fertigungsverfahren Wälzstoßen

Band 126
Dr.-Ing. Martin Gutmann

Entwicklung einer methodischen Vorgehensweise zur Diagnose von hydraulischen Produktionsmaschinen

Band 127
Dr.-Ing. Gisela Lanza

Simulative Anlaufunterstützung auf Basis der Qualitätsfähigkeiten von Produktionsprozessen

Band 128
Dr.-Ing. Ulf Dambacher

Kugelgewindetrieb mit hohem Druckwinkel

Band 129
Dr.-Ing. Carsten Buchholz

Systematische Konzeption und Aufbau einer automatisierten Produktionszelle für pulverspritzgegossene Mikrobauteile

Band 130
Dr.-Ing. Heiner Lang

Trocken-Räumen mit hohen Schnittgeschwindigkeiten

Band 131
Dr.-Ing. Daniel Nesges

Prognose operationeller Verfügbarkeiten von Werkzeugmaschinen unter Berücksichtigung von Serviceleistungen

Im Shaker Verlag erschienene Bände:

Band 132
Dr.-Ing. Andreas Bechle

Beitrag zur prozesssicheren Bearbeitung beim Hochleistungs-fertigungsverfahren Wälzschälen

Band 133
Dr.-Ing. Markus Herm

Konfiguration globaler Wertschöpfungsnetzwerke auf Basis von Business Capabilities

Band 134
Dr.-Ing. Hanno Tritschler

Werkzeug- und Zerspanprozessoptimierung beim Hartfräsen von Mikrostrukturen in Stahl

Band 135
Dr.-Ing. Christian Munzinger

Adaptronische Strebe zur Steifigkeitssteigerung von Werkzeugmaschinen

Band 136
Dr.-Ing. Andreas Stepping

Fabrikplanung im Umfeld von Wertschöpfungsnetzwerken und ganzheitlichen Produktionssystemen

Band 137
Dr.-Ing. Martin Dyck

Beitrag zur Analyse thermische bedingter Werkstückdeformationen in Trockenbearbeitungsprozessen

Band 138
Dr.-Ing. Siegfried Schmalzried

Dreidimensionales optisches Messsystem für eine effizientere geometrische Maschinenbeurteilung

Band 139
Dr.-Ing. Marc Wawerla

Risikomanagement von Garantieleistungen

Band 140
Dr.-Ing. Ivesa Buchholz

Strategien zur Qualitätssicherung mikromechanischer Bauteile mittels multisensorieller Koordinatenmesstechnik

Band 141
Dr.-Ing. Jan Kotschenreuther

Empirische Erweiterung von Modellen der Makrozerspanung auf den Bereich der Mikrobearbeitung

Band 142
Dr.-Ing. Andreas Knödel

Adaptronische hydrostatische Drucktascheneinheit

Band 143
Dr.-Ing. Gregor Stengel

Fliegendes Abtrennen räumlich gekrümmter Strangpressprofile mittels Industrierobotern

Band 144
Dr.-Ing. Udo Weismann

Lebenszyklusorientiertes interorganisationelles Anlagencontrolling

Band 145
Dr.-Ing. Rüdiger Pabst

Mathematische Modellierung der Wärmestromdichte zur Simulation des thermischen Bauteilverhaltens bei der Trockenbearbeitung

Band 146
Dr.-Ing. Jan Wieser

Intelligente Instandhaltung zur Verfügbarkeitssteigerung von Werkzeugmaschinen

Band 147
Dr.-Ing. Sebastian Haupt

Effiziente und kostenoptimale Herstellung von Mikrostrukturen durch eine Verfahrenskombination von Bahnerosion und Laserablation

Band 148
Dr.-Ing. Matthias Schlipf

Statistische Prozessregelung von Fertigungs- und Messprozess zur Erreichung einer variabilitätsarmen Produktion mikromechanischer Bauteile

Band 149
Dr.-Ing. Jan Philipp Schmidt-Ewig

Methodische Erarbeitung und Umsetzung eines neuartigen Maschinenkonzeptes zur produktflexiblen Bearbeitung räumlich gekrümmter Strangpressprofile

Band 150
Dr.-Ing. Thomas Ender

Prognose von Personalbedarfen im Produktionsanlauf unter Berücksichtigung dynamischer Planungsgrößen

Band 151
Dr.-Ing. Kathrin Peter

Bewertung und Optimierung der Effektivität von Lean Methoden in der Kleinserienproduktion

Band 152
Dr.-Ing. Matthias Schopp

Sensorbasierte Zustandsdiagnose und -prognose von Kugelgewindetrieben

Band 153
Dr.-Ing. Martin Kipfmüller

Aufwandsoptimierte Simulation von Werkzeugmaschinen

Band 154
Dr.-Ing. Carsten Schmidt

Development of a database to consider multi wear mechanisms within chip forming simulation

Band 155
Dr.-Ing. Stephan Niggeschmidt

Ausfallgerechte Ersatzteilbereitstellung im Maschinen- und Anlagenbau mittels lastabhängiger Lebensdauerprognose

Band 156
Dr.-Ing. Jochen Conrad Peters

Bewertung des Einflusses von Formabweichungen in der Mikro-Koordinatenmesstechnik

Band 157
Dr.-Ing. Jörg Ude

Entscheidungsunterstützung für die Konfiguration globaler Wertschöpfungsnetzwerke

Band 158
Dr.-Ing. Stefan Weiler

Strategien zur wirtschaftlichen Gestaltung der globalen Beschaffung

Band 159
Dr.-Ing. Jan Rühl

Monetäre Flexibilitäts- und Risikobewertung

Band 160
Dr.-Ing. Daniel Ruch

Positions- und Konturerfassung räumlich gekrümmter Profile auf Basis bauteilimmanenter Markierungen

Band 161
Dr.-Ing. Manuel Tröndle

Flexible Zuführung von Mikrobauteilen mit piezoelektrischen Schwingförderern

Band 162
Dr.-Ing. Benjamin Viering

Mikroverzahnungsnormal

Band 163
Dr.-Ing. Chris Becke

Prozesskraftrichtungsangepasste Frässtrategien zur schädigungsarmen Bohrungsbearbeitung an faserverstärkten Kunststoffen

Band 164
Dr.-Ing. Patrick Werner

Dynamische Optimierung und Unsicherheitsbewertung der lastabhängigen präventiven Instandhaltung von Maschinenkomponenten

Band 165
Dr.-Ing. Martin Weis

Kompensation systematischer Fehler bei Werkzeugmaschinen durch self-sensing Aktoren

Band 166
Dr.-Ing. Markus Schneider

Kompensation von Konturabweichungen bei gerundeten Strangpressprofilen durch robotergestützte Führungswerkzeuge

Band 167
Dr.-Ing. Ester M. R. Ruprecht

Prozesskette zur Herstellung schichtbasierter Systeme mit integrierten Kavitäten

Band 168
Dr.-Ing. Alexander Broos

Simulationsgestützte Ermittlung der Komponentenbelastung für die Lebensdauerprognose an Werkzeugmaschinen

Band 169
Dr.-Ing. Frederik Zanger

Segmentspanbildung, Werkzeugverschleiß, Randschichtzustand und Bauteileigenschaften: Numerische Analysen zur Optimierung des Zerspanungsprozesses am Beispiel von Ti-6Al-4V

Band 170
Dr.-Ing. Benjamin Behmann

Servicefähigkeit

Band 171
Dr.-Ing. Annabel Gabriele Jondral

Simulationsgestützte Optimierung und Wirtschaftlichkeitsbewertung des Lean-Methodeneinsatzes

Band 172
Dr.-Ing. Christoph Ruhs

Automatisierte Prozessabfolge zur qualitätssicheren Herstellung von Kavitäten mittels Mikrobahnerosion

Band 173
Dr.-Ing. Steven Peters

Markoffsche Entscheidungsprozesse zur Kapazitäts- und Investitionsplanung von Produktionssystemen

Band 174
Dr.-Ing. Christoph Kühlewein

Untersuchung und Optimierung des Wälzschälverfahrens mit Hilfe von 3D-FEM-Simulation – 3D-FEM Kinematik- und Spanbildungssimulation

Band 175
Dr.-Ing. Adam-Mwanga Dieckmann

Auslegung und Fertigungsprozessgestaltung sintergefügter Verbindungen für µMIM-Bauteile

Band 176
Dr.-Ing. Heiko Hennrich

Aufbau eines kombinierten belastungs- und zustandsorientierten Diagnose- und Prognosesystems für Kugelgewindetriebe

Band 177
Dr.-Ing. Stefan Herder

Piezoelektrischer Self-Sensing-Aktor zur Vorspannungsregelung in adaptronischen Kugelgewindetrieben

Band 178
Dr.-Ing. Alexander Ochs

Ultraschall-Strömungsgreifer für die Handhabung textiler Halbzeuge bei der automatisierten Fertigung von RTM-Bauteilen

Band 179
Dr.-Ing. Jürgen Michna

Numerische und experimentelle Untersuchung zerspanungsbedingter Gefügeumwandlungen und Modellierung des thermo-mechanischen Lastkollektivs beim Bohren von 42CrMo4

Band 180
Dr.-Ing. Jörg Elser

Vorrichtungsfreie räumliche Anordnung von Fügepartnern auf Basis von Bauteilmarkierungen

Band 181
Dr.-Ing. Katharina Klimscha

Einfluss des Fügespalts auf die erreichbare Verbindungsqualität beim Sinterfügen

Band 182
Dr.-Ing. Patricia Weber

Steigerung der Prozesswiederholbarkeit mittels Analyse akustischer Emissionen bei der Mikrolaserablation mit UV-Pikosekundenlasern

Band 183
Dr.-Ing. Jochen Schädel

Automatisiertes Fügen von Tragprofilen mittels Faserwickeln

Band 184
Dr.-Ing. Martin Krauße

Aufwandsoptimierte Simulation von Produktionsanlagen durch Vergrößerung der Geltungsbereiche von Teilmodellen

Band 185
Dr.-Ing. Raphael Moser

Strategische Planung globaler Produktionsnetzwerke
Bestimmung von Wandlungsbedarf und Wandlungszeitpunkt mittels multikriterieller Optimierung

Band 186
Dr.-Ing. Martin Otter

Methode zur Kompensation fertigungsbedingter Gestaltabweichungen für die Montage von Aluminium Space-Frame-Strukturen

Band 187
Dr.-Ing. Urs Leberle

Produktive und flexible Gleitförderung kleiner Bauteile auf phasenflexiblen Schwingförderern mit piezoelektrischen 2D-Antriebselementen

Band 188
Dr.-Ing. Johannes Book

Modellierung und Bewertung von Qualitätsmanagementstrategien in globalen Wertschöpfungsnetzwerken

Band 189
Dr.-Ing. Florian Ambrosy

Optimierung von Zerspanungsprozessen zur prozesssicheren Fertigung nanokristalliner Randschichten am Beispiel von 42CrMo4

Band 190
Dr.-Ing. Adrian Kölmel

Integrierte Messtechnik für Prozessketten unreifer Technologien am Beispiel der Batterieproduktion für Elektrofahrzeuge

Band 191
Dr.-Ing. Henning Wagner

Featurebasierte Technologieplanung zum Preforming von textilen Halbzeugen

Band 192
Dr.-Ing. Johannes Gebhardt

Strukturoptimierung von in FVK eingebetteten metallischen Lasteinleitungselementen

Band 193
Dr.-Ing. Jörg Bauer

Hochintegriertes hydraulisches Vorschubsystem für die Bearbeitung kleiner Werkstücke mit hohen Fertigungsanforderungen

Band 194
Dr.-Ing. Nicole Stricker

Robustheit verketteter Produktionssysteme
Robustheitsevaluation und Selektion des Kennzahlensystems der Robustheit

Band 195
Dr.-Ing. Anna Sauer

Konfiguration von Montagelinien unreifer Produkttechnologien am Beispiel der Batteriemontage für Elektrofahrzeuge

Band 196
Dr.-Ing. Florian Sell-Le Blanc

Prozessmodell für das Linearwickeln unrunder Zahnspulen
Ein Beitrag zur orthozyklischen Spulenwickeltechnik

Band 197
Dr.-Ing. Frederic Förster

Geregeltes Handhabungssystem zum zuverlässigen und energieeffizienten Handling textiler Kohlenstofffaserzuschnitte

Band 198
Dr.-Ing. Nikolay Boev

Numerische Beschreibung von Wechselwirkungen zwischen Zerspanprozess und Maschine am Beispiel Räumen

Band 199
Dr.-Ing. Sebastian Greinacher

Simulationsgestützte Mehrzieloptimierung schlanker und ressourceneffizienter Produktionssysteme

Band 200
Dr.-Ing. Benjamin Häfner

Lebensdauerprognose in Abhängigkeit der Fertigungsabweichungen bei Mikroverzahnungen

Band 201
Dr.-Ing. Stefan Klotz

Dynamische Parameteranpassung bei der Bohrungsherstellung in faserverstärkten Kunststoffen unter zusätzlicher Berücksichtigung der Einspannsituation

Band 202
Dr.-Ing. Johannes Stoll

Bewertung konkurrierender Fertigungsfolgen mittels Kostensimulation und stochastischer Mehrzieloptimierung
Anwendung am Beispiel der Blechpaketfertigung für automobile Elektromotoren

Band 203
Dr.-Ing. Simon-Frederik Koch

Fügen von Metall-Faserverbund-Hybridwellen im Schleuderverfahren
ein Beitrag zur fertigungsgerechten intrinsischen Hybridisierung

Band 204
Dr.-Ing. Julius Ficht

Numerische Untersuchung der Eigenspannungsentwicklung für sequenzielle Zerspanungsprozesse

Band 205
Dr.-Ing. Manuel Baumeister

Automatisierte Fertigung von Einzelblattstapeln in der Lithium-Ionen-Zellproduktion

Band 206
Dr.-Ing. Daniel Bertsch

Optimierung der Werkzeug- und Prozessauslegung für das Wälzschälen von Innenverzahnungen

Band 207
Dr.-Ing. Kyle James Kippenbrock

Deconvolution of Industrial Measurement and Manufacturing Processes for Improved Process Capability Assessments

Band 208
Dr.-Ing. Farboud Bejnoud

Experimentelle Prozesskettenbetrachtung für Räumbauteile am Beispiel einer einsatzgehärteten PKW-Schiebemuffe

Band 209
Dr.-Ing. Steffen Dosch

Herstellungsübergreifende Informationsübertragung zur effizienten Produktion von Werkzeugmaschinen am Beispiel von Kugelgewindetrieben

Band 210
Dr.-Ing. Emanuel Moser

Migrationsplanung globaler Produktionsnetzwerke
Bestimmung robuster Migrationspfade und risiko-effizienter Wandlungsbefähiger

Band 211
Dr.-Ing. Jan Hochdörffer

Integrierte Produktallokationsstrategie und Konfigurationssequenz in globalen Produktionsnetzwerken

Band 212
Dr.-Ing. Tobias Arndt

Bewertung und Steigerung der Prozessqualität in globalen Produktionsnetzwerken

Band 213
Dr.-Ing. Manuel Peter

Unwuchtminimale Montage von Permanentmagnetrotoren durch modellbasierte Online-Optimierung

Band 214
Dr.-Ing. Robin Kopf

Kostenorientierte Planung von Fertigungsfolgen additiver Technologien

Band 215
Dr.-Ing. Harald Meier

Einfluss des Räumens auf den Bauteilzustand in der Prozesskette Weichbearbeitung – Wärmebehandllung – Hartbearbeitung

Band 216
Dr.-Ing. Daniel Brabandt

Qualitätssicherung von textilen Kohlenstofffaser-Preforms mittels optischer Messtechnik

Band 217
Dr.-Ing. Alexandra Schabunow

Einstellung von Aufnahmeparametern mittels projektionsbasierter Qualitäts-kenngrößen in der industriellen Röntgen-Computertomographie

Band 218
Dr.-Ing. Jens Bürgin

Robuste Auftragsplanung in Produktionsnetzwerken
Mittelfristige Planung der variantenreichen Serienproduktion unter Unsicherheit der Kundenauftragskonfigurationen

Band 219
Dr.-Ing. Michael Gerstenmeyer

Entwicklung und Analyse eines mechanischen Oberflächenbehandlungs-verfahrens unter Verwendung des Zerspanungswerkzeuges

Band 220
Dr.-Ing. Jacques Burtscher

Erhöhung der Bearbeitungsstabilität von Werkzeugmaschinen durch semi-passive masseneinstellbare Dämpfungssysteme

Band 221
Dr.-Ing. Dietrich Berger

Qualitätssicherung von textilen Kohlenstofffaser-Preforms mittels prozess-integrierter Wirbelstromsensor-Arrays

Band 222
Dr.-Ing. Fabian Johannes Ballier

Systematic gripper arrangement for a handling device in lightweight production processes

Band 223
Dr.-Ing. Marielouise Schäferling, geb. Zaiß

Development of a Data Fusion-Based Multi-Sensor System for Hybrid Sheet Molding Compound

Band 224
Dr.-Ing. Quirin Spiller

Additive Herstellung von Metallbauteilen mit dem ARBURG Kunststoff-Freiformen

Band 225
Dr.-Ing. Andreas Spohrer

Steigerung der Ressourceneffizienz und Verfügbarkeit von Kugelgewindetrieben durch adaptive Schmierung

Band 226
Dr.-Ing. Johannes Fisel

Veränderungsfähigkeit getakteter Fließmontagesysteme
Planung der Fließbandabstimmung am Beispiel der Automobilmontage

Band 227
Dr.-Ing. Patrick Bollig

Numerische Entwicklung von Strategien zur Kompensation thermisch bedingter Verzüge beim Bohren von 42CrMo4

Band 228
Dr.-Ing. Ramona Pfeiffer, geb. Singer

Untersuchung der prozessbestimmenden Größen für die anforderungsgerechte Gestaltung von Pouchzellen-Verpackungen

Band 229
Dr.-Ing. Florian Baumann

Additive Fertigung von endlosfaserverstärkten Kunststoffen mit dem ARBURG Kunststoff-Freiform Verfahren

Band 230
Dr.-Ing. Tom Stähr

Methodik zur Planung und Konfigurationsauswahl skalierbarer Montagesysteme – Ein Beitrag zur skalierbaren Automatisierung

Band 231
Dr.-Ing. Jan Schwennen

Einbringung und Gestaltung von Lasteinleitungsstrukturen für im RTM-Verfahren hergestellte FVK-Sandwichbauteile

Band 232
Dr.-Ing. Sven Coutandin

Prozessstrategien für das automatisierte Preforming von bebinderten textilen Halbzeugen mit einem segmentierten Werkzeugsystem

Band 233
Dr.-Ing. Christoph Liebrecht

Entscheidungsunterstützung für den Industrie 4.0-Methodeneinsatz
Strukturierung, Bewertung und Ableitung von Implementierungsreihenfolgen

Band 234
Dr.-Ing. Stefan Treber

Transparenzsteigerung in Produktionsnetzwerken
Verbesserung des Störungsmanagements durch verstärkten Informationsaustausch

Band 235
Dr.-Ing. Marius Dackweiler

Modellierung des Fügewickelprozesses zur Herstellung von leichten Fachwerkstrukturen

Band 236
Dr.-Ing. Fabio Echsler Minguillon

Prädiktiv-reaktives Scheduling zur Steigerung der Robustheit in der Matrix-Produktion

Band 237
Dr.-Ing. Sebastian Haag

Entwicklung eines Verfahrensablaufes zur Herstellung von Batteriezellstapeln mit großformatigem, rechteckigem Stapelformat und kontinuierlichen Materialbahnen

Band 238
Dr.-Ing. Raphael Wagner

Strategien zur funktionsorientierten Qualitätsregelung in der Serienproduktion

Band 239
Dr.-Ing. Christopher Ehrmann

Ausfallfrüherkennung von Ritzel-Zahnstangen- Trieben mittels Acoustic Emission

Band 240
Dr.-Ing. Janna Hofmann

Prozessmodellierung des Fünf-Achs-Nadelwickelns zur Implementierung einer trajektoriebasierten Drahtzugkraftregelung

Band 241
Dr.-Ing. Andreas Kuhnle

Adaptive Order Dispatching based on Reinforcement Learning
Application in a Complex Job Shop in the Semiconductor Industry

Band 242
Dr.-Ing. Andreas Greiber

Fertigung optimierter technischer Oberflächen durch eine Verfahrenskombination aus Fliehkraft-Tauchgleitschleifen und Laserablation
Prozesseinflüsse und Prozessauslegung

Band 243
Dr.-Ing. Jan Niclas Eschner

Entwicklung einer akustischen Prozessüberwachung zur Porenbestimmung im Laserstrahlschmelzen

Band 244
Dr.-Ing. Sven Roth

Schädigungsfreie Anbindung von hybriden FVK/Metall-Bauteilen an metallische Tragstrukturen durch Widerstandspunktschweißen

Band 245
Dr.-Ing. Sina Kathrin Peukert

Robustheitssteigerung in Produktionsnetzwerken mithilfe eines integrierten Störungsmanagements

Band 246
Dr.-Ing. Alexander Jacob

Hochiterative Technologieplanung
Rekursive Optimierung produkt- und fertigungsbezogener Freiheitsgrade am Beispiel der hybrid-additiven Fertigung

Band 247
Dr.-Ing. Patrick Moll

Ressourceneffiziente Herstellung von Langfaser-Preforms im Faserblasverfahren

Band 248
Dr.-Ing. Eric Thore Segebade

Erhöhung der Verschleißbeständigkeit von Bauteilen aus Ti-6Al-4V mittels simulationsgestützter Zerspanung und mechanischer Mikrotexturierung

Band 249
Dr.-Ing. Shun Yang

Regionalized implementation strategy of smart automation within assembly systems in China

Band 250
Dr.-Ing. Constantin Carl Hofmann

Vorausschauende und reaktive Mehrzieloptimierung für die Produktionssteuerung einer Matrixproduktion

Band 251
Dr.-Ing. Paul Ruhland

Prozesskette zur Herstellung von hybriden Faser-Metall-Preforms
Modellbildung und Optimierung des Binderauftrags und der Drapierung für stabförmige Bauteile

Band 252
Dr.-Ing. Leonard Schild

Erzeugung und Verwendung von Anwendungswissen in der industriellen Computertomographie

Band 253
Dr.-Ing. Benedikt Klee

Analyse von Phaseninformationen in Videodaten zur Identifikation von Schwingungen in Werkzeugmaschinen

Band 254
Dr.-Ing. Bruno Vargas

Wälzschälen mit kleinen Achskreuzwinkeln
Prozessgrenzen und Umsetzbarkeit

Band 255
Dr.-Ing. Lucas Bretz

Function-oriented in-line quality assurance of hybrid sheet molding compound

Band 256
Dr.-Ing. Bastian Rothaupt

Dämpfung von Bauteilschwingungen durch einstellbare Werkstückdirektspannung mit Hydrodehnspanntechnik

Band 257
Dr.-Ing. Daniel Kupzik

Robotic Swing Folding of three-dimensional UD-tape-based Reinforcement Structures

Band 258
Dr.-Ing. Bastian Verhaelen

(De-)Zentralisierung von Entscheidungen in globalen Produktionsnetzwerken
Strategie- und komplexitätsorientierte Gestaltung der Entscheidungsautonomie

Band 259
Dr.-Ing. Hannes Wilhelm Weinmann

Integration des Vereinzelungs- und Stapelbildungsprozesses in ein flexibel und kontinuierlich arbeitendes Anlagenmodul für die Li-Ionen Batteriezellfertigung

Band 260
Dr.-Ing. Florian Stamer

Dynamische Lieferzeit-Preisgestaltung in variantenreicher Produktion
Ein adaptiver Ansatz mithilfe von Reinforcement Learning

Band 261
Dr.-Ing. Patrick Neuenfeldt

Modellbildung des Tauchgleitschleifens zur Abtrag- und Topografievorhersage an komplexen Geometrien

Band 262
Dr.-Ing. Boris Matuschka

Energieeffizienz in Prozessketten: Analyse und Optimierung von Energieflüssen bei der Herstellung eines PKW-Getriebebauteils aus 16MnCr5

Band 263
Dr.-Ing. Tobias Schlagenhauf

Bildbasierte Quantifizierung und Prognose des Verschleißes an Kugelgewindetriebspindeln
Ein Beitrag zur Zustandsüberwachung von Kugelgewindetrieben mittels Methoden des maschinellen Lernens

Band 264
Dr.-Ing. Benedict Stampfer

Entwicklung eines multimodalen Prozessmodells zur Oberflächenkonditionierung beim Außenlängsdrehen von 42CrMo4

Band 265
Dr.-Ing. Carmen Maria Krahe

KI-gestützte produktionsgerechte Produktentwicklung
Automatisierte Wissensextraktion aus vorhandenen Produktgenerationen

Band 266
Dr.-Ing. Markus Netzer

Intelligente Anomalieerkennung für hochflexible Produktionsmaschinen
Prozessüberwachung in der Brownfield Produktion

Band 267
Dr.-Ing. Simon Raphael Merz

Analyse der Kinematik und Kinetik von Planetenwälzgewindetrieben

Band 268
Dr.-Ing. Rainer Maria Silbernagel

Funktionsorientierte Qualitätsregelung in Produktionsnetzwerken
Qualitätsmanagement in der Produktion hochpräziser Produkte durch netzwerkweite Datenintegration

Band 269
Dr.-Ing. Jonas Nieschlag

Gestaltung und Prozessanalyse für im Schleuderverfahren hergestellte FKV-Metall-Hohlstrukturen

Band 270
Dr.-Ing. Lukas Matthias Weiser

In-Process Porositätserkennung für den PBF-LB/M-Prozess

Band 271
Dr.-Ing. Leonard Vincent Overbeck

Digital Twins of production systems
Automated validation and update of material flow simulation models with real data

Band 272
Dr.-Ing. Felix Klenk

Transparenzsteigerung in der Rückführungslogistik zur Verbesserung der Materialbedarfsplanung für das Remanufacturing

Band 273
Dr.-Ing. Benjamin Bold

Kompensation der Wrinkle-Bildung beim Kalandrieren von Lithium-Ionen-Kathoden
Vom Prozessverständnis des Kalandrierens bis zur Prozessoptimierung mittels Anti-Wrinkle-Modul

Band 274
Dr.-Ing. Daniel Gauder

Adaptive in-line Qualitätsregelung in der Mikro-Verzahnungsfertigung

Band 275
Dr.-Ing. Fabian Sasse

Ontologie-basierte Entscheidungsunterstützung für die Auswahl von Messsystemen in unreifen Produktionsprozessen

Band 276
Dr.-Ing. Jonas Hillenbrand

Unsupervised Condition-Monitoring für Kugelgewindetriebe mittels Acoustic Emission

Band 277
Dr.-Ing. Manuela Neuenfeldt

Untersuchung des Einflusses der PBF-LB-Stellgrößen auf die zerspanende Bearbeitung additiv gefertigter Stahlbauteile

Band 278
Dr.-Ing. Marvin Carl May

Intelligent production control for time-constrained complex job shops