Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

CONTINUOUS FLOW TRANSPORT SCHEDULING FOR CONVEYOR-BASED AMHS IN WAFER FABS

Clemens Schwenke
Klaus Kabitzsch

Department of Applied Computer Science
Dresden University of Technology (TUD)
01062 Dresden, GERMANY

ABSTRACT

Automated material handling systems (AMHS) can greatly impact the manufacturing performance of a semiconductor fabricating facility (fab). High traffic loads within an AMHS can impede individual wafer lots so that they arrive late at their destination machines. Thus, corresponding process operations as well as dependent succeeding operations will be delayed due to the fab schedule's precedence constraints. Consequently, such transport-related delays can widely propagate throughout the overall fab schedule. In order to reduce transport-related delays before time-critical operations, novel ways of planning wafer transports have been investigated in this study. For validation, a well-known realistic representative wafer fab model has been extended with conveyor elements constituting a typical AMHS for continuous flow transport (CFT). As a result, improvements of the overall fab performance due to advanced transport scheduling methods are demonstrated and compared. Finally, the practicality of the suggested methods is discussed in the dynamic scheduling context of real fabs.

1 INTRODUCTION

Modern semiconductor fabricating facilities exhibit complex control systems for planning and scheduling the production of wafers. While planning of production and order releases is done at the top of the control hierarchy, scheduling of process operations on machines is done at the lower fab scheduling level. On the subordinate transportation level, the AMHS transfers the wafer lots from one machine to another (Mönch et al. 2012).

In order to perform the production, the controls on both lower levels have to work together. That is, for any machine that turned idle, the fab scheduling decides which wafer lot shall be processed next. And the AMHS carries this wafer lot to the location of that idle machine. In real wafer fabs, schedulers often assume average transport times which they gathered from observations or time stamped recordings. But unfortunately, real AMHS do not always deliver the wafer lots in the expected time. Instead, transport delays can be observed (Temponi et al. 2012).

These transport delays can be caused by wafer lots impeding each other and forming queues on conveyors during periods of high traffic loads. Consequently, the interaction of fab scheduling and transport scheduling should be investigated and improved (Mönch et al. 2011). Accordingly, this paper suggests transport scheduling methods for a better cooperation between the two control levels and compares them to a state-of-the-art control for conveyor-based AMHS. Hence, the overall objective of this study is a better integration of fab scheduling and transport scheduling that can be easily installed in existing fabs.

The latter of this paper is structured as follows. Related work is described in Section 2. In Section 3, the overall approach of this study is explained. Section 4 contains a validation using wafer fab benchmark data (Fowler and Robinson 1995) which we extended with an AMHS model. Finally, a conclusion is given discussing the approach's usability in the context of continuous dynamic scheduling in real fabs.

978-1-5386-3428-8/17/$31.00 ©2017 IEEE

2 STATE OF THE ART

In state of the art wafer fabs, resource conflicts occur within the machine setting as well as in the transport system (Mönch et al. 2012). Therefore, example work is reviewed that regards AMHS in some way either on the superordinate fab scheduling level or on the subordinate wafer lot transportation level.

2.1 Fab Scheduling

Commonly, practitioners produce their fab schedules, also named as machine (operation) schedules, by simulation of combined dispatching rules such as shortest processing time (SPT) or earliest due date (EDD) (Scholl et al. 2011). If more optimized fab schedules are desired, decomposition techniques (Sourirajan and Uzsoy 2007), shifting bottleneck approaches (Mason et al. 2002) or metaheuristic searches (Wang et al. 2013) have to be applied due to the large scale of a realistic fab scheduling problem. Alternatively, decomposed mixed-integer-programming-based approaches have been investigated (Klemmt et al. 2009).

Most of these approaches assume unlimited transport capacities and given average transport times. In contrast to these assumptions, few approaches exist which integrate the limited AMHS capacities into the overall fab scheduling problem (Qu et al. 2003, Deroussi et al. 2008, Drießel and Mönch 2012, Poppenborg et al. 2012, Zhang et al. 2012, Lacomme et al. 2013). But in these approaches the considered AMHS consists of either robots or vehicles. To the best knowledge of the authors there are no integrated scheduling approaches that consider conveyor-based AMHS in the context of semiconductor manufacturing.

2.2 Wafer Lot Transportation

Overviews of AMHS layouts and technologies are provided by Agrawal and Heragu (2006) and Montoya-Torres (2006). As Temponi et al. (2012) state, vehicle-based AMHS are very costly. Thus, the AMHS of next generation wafer fabs shall exhibit increasing capacity (Pettinato and Pillai 2005) and lower variability of delivery times but incur lower costs. Even before the 200mm-to-300mm transition, when still human operators moved the wafers inside many 200mm fabs, the benefits of low cost conveyor-based AMHS had been anticipated by Arzt and Bulcke (1999) and Brain et al. (1999). Arzt and Bulcke (1999) argued that vehicle-based AMHS often exceed investment costs of 30 million $US. Assuming 20 000 $US per one meter conveyor hardware and software, e.g., an AMHS of 183 meters would cost only 3 660 000 $US.

Accordingly, the potential of continuous wafer lot transportation has been successfully demonstrated by pioneers as Heinrich and Pyke (1999) throughout many years in one of the world's most highly automated fabs for 200mm wafers (Heinrich et al. 2008, Bannert et al. 2012). Hence, due to increasing cost pressure, the potential of conveyors may be (re)discovered by decision makers in the near future (Wang et al. 2016). Accordingly, we focus on conveyor-based as in contrast to vehicle-based AMHS.

At the superordinate fab scheduling level, the machines are limited resources and a scheduler assigns waiting wafers to idle machines. Similarly, this is done for the resource conflicts at the subordinate transport scheduling level. In a conveyor-based AMHS, the rotary tables at junctions are the limited resources (Hong et al. 2011). For the assessment of AMHS, on the one hand, there is work on approximate simulation (Jimenez et al. 2008, Hammel et al. 2012). Essentially, such methods identify overloaded track segments in order to either redesign the layout or to balance traffic loads (Zhang et al. 2016, Zhou et al. 2016). On the other hand, there exists specific work on detailed design and simulation of conveyor-based AMHS wherein the congestion effects of layout decisions are analyzed (Paprotny et al. 2000, Nazzal et al. 2010, Lasrado and Nazzal 2011). Most of this work is based on given wafer traffic flows that have to be processed by the rotary tables in a first-in-first-out (FIFO) manner, regardless of possible tailbacks.

Hence, Jiong et al. (2013) conclude that many CFT-based systems exhibit rather myopic controls and call for optimization of the AMHS to enable conflict-free movement. Consequently, our study aims at novel transport scheduling methods that guarantee conflict-free movement *a priori* for any given fab schedule or any given fab layout, cf. Subsections 3.3 and 3.4.

3 APPROACH

Before we were able to test transport scheduling methods, we had to create a suitable AMHS model, as described in Subsection 3.1. Afterwards, in order to compare three different transport scheduling methods, first an initial machine schedule, more specifically a machine operation sequence, had to be obtained for the fab, see Subsection 3.2. This initial machine schedule then was used as a basis for investigating three different transport scheduling methods, see Subsection 3.3. Finally, the resulting delays of process operations had to be converted into fab performance indicators such as job cycle time or job tardiness. As a result, the three transport scheduling methods were compared as reported in Subsection 3.4. In the following, the approach's four steps are described in greater detail, also see Figure 1.

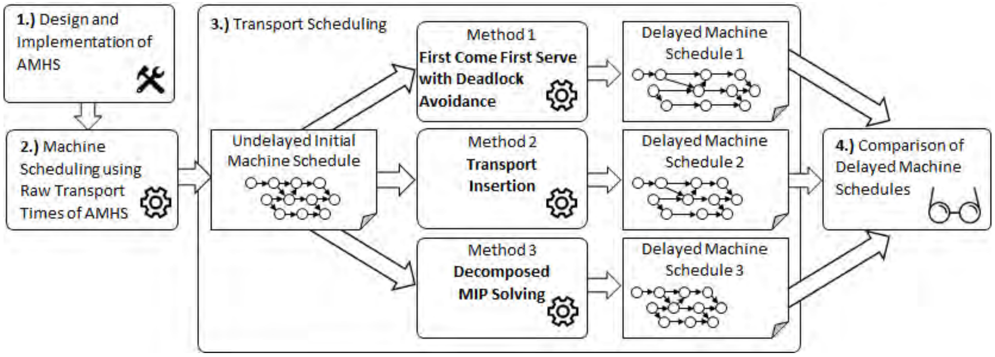
1.) Design and Implementation of AMHS
2.) Machine Scheduling using Raw Transport Times of AMHS
3.) Transport Scheduling
Undelayed Initial Machine Schedule
Method 1 First Come First Serve with Deadlock Avoidance
Method 2 Transport Insertion
Method 3 Decomposed MIP Solving
Delayed Machine Schedule 1
Delayed Machine Schedule 2
Delayed Machine Schedule 3
4.) Comparison of Delayed Machine Schedules

Figure 1: Overall workflow of approach.

3.1 Wafer Fab Model and Design of Suitable AMHS

For validating fab simulation or scheduling approaches there exist well-known reference (simulation) models of wafer fabs. These models originate from the Measurement and Improvement of Manufacturing Capacity (MIMAC) project and are referred to as MIMAC models (Fowler and Robinson 1995). These MIMAC models are widely used by researchers, which work on semiconductor industry problems, because they represent the typical settings and behaviors of wafer production very well. Back then, when the MIMAC models were developed, commonly human operators manually carried the wafers from one machine to another. Accordingly, the MIMAC models do not provide AMHS model components. As a result, we developed an AMHS model that is close to reality and fits to the MIMAC fab model SET2. This AMHS model emulates a continuous flow transport system consisting of 122 rotary tables and approx. 1162 meters of conveyors. The conveyor speed is 8.58 inches per second. The rotary tables' size is 12 inches. Traversing takes 6 seconds straight or 9 seconds in case of turning.

The layout design of this AMHS model followed three conflicting objectives. First, the transport system elements should be arranged in a spine-like interbay-bay layout that would be as realistic as possible. In Figure 2, the central horizontal conveyor double track models the inter-bay. The vertical branches depict the intra-bay conveyor tracks. Second, the locations of the rotary tables and the distances of the conveyors should be arranged in such a way that they would mimic the given transport times of MIMAC data SET2 nearly optimally. Hence, we pooled the machines, in the MIMAC data set named as tools, according to process operation type and located them in dedicated areas, called bays. Sometimes, shortcuts and further load port branches in front of machines were necessary to add distance in order to resemble the given transport times of MIMAC data SET2. Consequently, some shortcuts connect certain areas directly, as this can be seen in mature wafer fabs. In fact, practitioners often quickly install new conveyor segments to connect newly purchased machines with the existing (real) AMHS or sometimes also to relieve traffic loads from junctions at main thoroughfares. Third, the layout should contain as few conveyor elements and rotary tables as possible. As a result, a big portion of the traffic load and thus noticeable delays would occur at junctions in the inter-bay or at intersections near frequently visited machines or bottle neck tools.

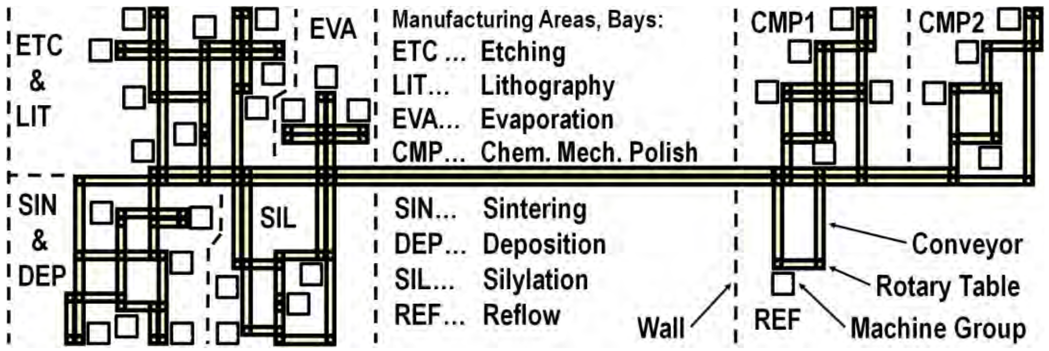

Figure 2: Conveyor-based AMHS model with horizontal interbay, vertical bays, various shortcuts and load port branches linking to machines for similarity with given transport times.

3.2 Initial Machine Schedule

In order to produce an optimal initial machine schedule, a flexible job shop scheduling problem (FJSSP) would have to be solved. According to Graham et al. (1979) it can be stated as $FJ_c\,|r_i,\ rcrc|\,TWT$. The jobs J_i enter at release dates r_i and exit at completion times C_i preferably before due dates d_i with a weight w_i. The jobs visit m machines, which are grouped in c workcenters, multiple times so that the jobs are recirculated ($rcrc$). The objective would be to minimize the total weighted tardiness $TWT = \sum_{i=1}^{n} w_i T_i$ (1). Tardiness is $T_i = \max(C_i - d_i, 0)$. The FJSSP was reduced to a JSSP by assigning the operations equally to suitable machines for load balancing. The resulting JSSP can be stated as follows.

$$\begin{array}{lllr} \text{Minimize} & TWT & & (1) \\ \text{s. t.} & S_{io} + p_{io} + t_{io} \le S_{jp} & O_{io}, O_{jp} \in V, \left(O_{io}, O_{jp}\right) \in C & (2) \\ & S_{io} + p_{io} \le S_{jp} \;||\; S_{jp} + p_{jp} \le S_{io} & O_{io}, O_{jp} \in V, \left(O_{io}, O_{jp}\right) \in D & (3) \end{array}$$

This formulation references the disjunctive graph $G := (V, C, D)$. The set of nodes V represents the process operations O_{io} of all jobs J_i. A job J_i consists of n_i operations $O_{io}$with precedence constraints $O_{i1} \rightarrow O_{i2} \rightarrow \cdots \rightarrow O_{in_i}$ of its process route. Each operation O_{io} is associated with a starting time S_{io} and process duration p_{io}. The set of edges C represents the process-related precedence constraints between the operations O_{io} of a job J_i. The set D represent the undirected edges modeling the capacity constraints of the machines. Thus, the precedence constraints (2) ensure the process operation sequence, given by each job's process route. The capacity constraints (3) ensure that each machine processes only one operation at a time.

For the production of this initial machine schedule the transport times t_{io} between one operation O_{io} and another O_{jp} are considered ideally short and static. Thus, transport operation conflicts are ignored. Such an ideally short transport time is computed by summing up all traversing times q_{kt} over each AMHS element t on the path connecting source machine and target machine for a transport job J'_k. Hence, an ideally short travel time represents the time it would take for a wafer lot to travel alone on its path, unhindered by any other wafer lots. Analogous to the fab scheduling term *raw processing time* $p_i = \sum_{o=1}^{n_i} p_{io}$ of one wafer production job J_i, in the context of transport scheduling this ideally short travel time can be called *raw transport time* $t_{io} = \sum_{t=1}^{n_k} q_{kt}$ for a transport job J'_k between two subsequent process operations O_{io} and O_{jp}.

For quickly producing an idealized initial machine schedule regardless of its optimality, a simple forward simulation was performed. Thus, one of the dispatching rules FIFO, SPT or EDD was applied whenever a machine became idle and needed to have a new job assigned. In summary, from a simulation point of view, the initial schedule was constructed by simulation of dispatching rules. By doing so, from a scheduling point of view, the disjunctive edges of the underlying disjunctive graph G have been oriented

(Figure 3, dashed arrow lines between round nodes). And in order to maintain comparability, these orientations shall not be changed anymore for any further investigation of the different transport scheduling methods. Thus, favoring the comparability, the superordinate machine schedule is fixed and not optimal. Instead, only the subordinate transport schedule shall be optimized during the further procedure.

3.3 Transport Scheduling Methods

Three different methods for scheduling the transports have been investigated. The first method resembles a common transport system logic as it can be found in many real CFT-based AMHS. It is called *First Come First Served with Deadlock Avoidance* and serves as the benchmark for the following two other novel methods. The second method successively fills a transport schedule and is called *Transport Insertion* method. The third method decomposes the underlying overall transport scheduling problem into small mixed integer problems (MIP) of possibly interfering transports. Then it solves these MIPs optimally by exploiting slacks of the initial machine schedule. It is called *Decomposed MIP Solving* method or simply *MIP-based* method. Each method uses the formerly produced initial machine schedule to derive the starting times of the relevant transport jobs that shall be simulated or scheduled next.

Importantly, after each transport, that is found to be delayed, this transport-related delay has to be fed back into the initial machine schedule. As a result, succeeding machine operations after this delayed transport will be delayed as well. This feedback mechanism is crucial because the delayed target machine operations in turn cause their subsequent transports to start delayed too. Accordingly, further dependent succeeding machine operations will be delayed as well as their subsequent transports and so forth. In short, after one transport has been delayed, the initial machine schedule is not valid anymore. Specifically, the starting times of subsequent machine operations and transport jobs are incorrect.

As a result, these starting times need to be updated. For simplicity and certainty the complete machine schedule could be updated using the critical path method (CPM). But in order to speed up this iterative update and feedback process, a reduced method has been implemented, which first identifies the impacted target machine operations and then only updates those subsequent transport jobs that will be simulated or scheduled next. This way, a large portion of the machine schedule remains not updated and all the computing effort for updating machine operations that are very far in the future, and therefore would be updated many times before their transport jobs actually need to be simulated or scheduled, can be saved.

3.3.1 First Come First Served with Deadlock Avoidance

The first method resembles the priority rules at conveyor intersections as they are in use in most real wafer fabs that have a CFT-based AMHS. At each rotary table the wafer lot, which arrived first, will have the right of way. Furthermore, at intersections a reservation mechanism avoids deadlocks, for example of four wafer lots each waiting for another one to get out of the way. This method is implemented as a deterministic discrete event simulation without any random elements. The initial machine schedule serves as input for the transports to be simulated. As noted, transport-related delays are fed back into the initial machine schedule in order to update the next transport's starting time.

3.3.2 Transport Insertion

The transport jobs J'_k are modeled similar to the process jobs J_i. Each transport job J'_k consists of n_k transport operations T_{kt} with precedence constraints $T_{k1} \rightarrow T_{k2} \rightarrow \cdots \rightarrow T_{kn_k}$ of its transport route between two subsequent process job operations $O_{io}, O_{jp} \in V, (O_{io}, O_{jp}) \in C$. Each transport operation T_{kt} is associated with a starting time S_{kt} and transfer duration q_{kt} over a rotary table or conveyor. The objective would be to minimize the total weighted completion time $TWC = \sum_{k=1}^{l} w_k C_k$ (4), complementary to minimizing TWT (1) on the fab scheduling level. The set of nodes W represents the process operations T_{kt} of the transport jobs J'_k. The weight w_k of a transport job represents its priority or urgency. The set of edges

E represents the path-related precedence constraints between the operations T_{kt} of a transport job J'_k. The set F represents the undirected edges modeling the capacity constraints of the rotary tables. Defining a disjunctive graph $H := (W, E, F, P)$, the transport scheduling problem can be stated as follows.

$$\begin{array}{llll}
\text{Minimize} & TWC & & (4)\\
\text{s. t.} & S_{kt} + q_{kt} = S_{lu} & T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in E & (5)\\
 & S_{kt} + q_{kt} \leq S_{lu} \;||\; S_{lu} + q_{lu} \leq S_{kt} & T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in F & (6)\\
 & S_{kt} + q_{kt} + p_{i\,o+1} \leq S_{lu} & T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in P & (7)
\end{array}$$

The precedence constraints (5) ensure the transport operation sequence of each transport job's travel route. Note that constraints (5) demand that one transport element is immediately entered after exiting the previous. Hence, *no-wait* constraints (nwt) are introduced. The capacity constraints (6) ensure that on each rotary table there is only one transport job lot J'_k at a time. Accordingly, set F only contains edges from one rotary table to another because interjacent conveyors can hold more than one wafer lot J'_k. Additionally, constraints (7) ensure the order of transport jobs. The order of transport jobs is given by the process-related order of corresponding process operations, e. g., $O_{io} \rightarrow O_{i\,o+1} \rightarrow O_{i\,o+2}$. Correspondingly, set P contains edges between the last operation T_{kt} of a transport job J'_k before process operation $O_{i\,o+1}$ and the first operation T_{lu} of a subsequent transport job J'_l after process operation $O_{i\,o+1}$. Hence, precedence constraints ($prec$) are introduced. In Graham's notation (Graham et al. 1979) the problem can be stated as $J'_b\,|r_k,\ prec,\ nwt|\,TWC$. The transport jobs J'_k have to traverse at most over b AMHS elements. A transport job J'_k enters no earlier than at a release date r_k given by the end of its preceding process operation O_{io}. Respectively, a subsequent transport job J'_l of set P enters at release date r_l, given by the end of $O_{i\,o+1}$.

The method of transport insertion proceeds as follows. First, the transport jobs are ordered by their earliest possible starting times r_k. Then, for each transport J'_k a coherent time window over all path elements between source and target machine is sought so that the transport operations $T_{k1} \rightarrow T_{k2} \rightarrow \cdots \rightarrow T_{kn_j}$ can be carried out unimpeded. If the first wide enough time window is found, it will be occupied. Precisely, the transport starting time S_{k1} will be set to the beginning to this time window and the transport will be inserted into the overall transport schedule. Hence, the starting times S_{kt} of the subsequent transport operations T_{kt} will be delayed the same amount as T_{k1}. This delay is $d_k = S_{k1} - r_k$. As a result, the transports start delayed but then travel undelayed throughout the AMHS model. That is, a wafer lot does not wait at rotary tables until another wafer lot exits. Instead, the wafer lot arrives no earlier than a previous one just left.

As a result, waiting queues or "traffic jams" will not form and classic event-based simulation is not needed. Instead, a transport schedule is constructed, or rather filled, by successively inserting transports. After delaying and inserting a transport, the corresponding target machine operation as well as dependent succeeding operations have to be delayed and updated in the initial machine schedule. Thus, the effects of the transport delays are successively fed back into the initial machine schedule.

3.3.3 Decomposed MIP Solving

The third method aims at minimizing the transport-delays which delay the overall machine schedule. For this purpose, at first the CPM is applied to the initial machine schedule. Performing a forward and backward walk, the CPM yields time windows for each machine operation. Such a time window not only provides the time for when an operation O_{io} can start the earliest but also for when it should end the latest in order to not delay the overall schedule. These windows are also known as slack or float. These slacks are exploited to set the weights w_k for the objective of minimizing TWC (4). Hence, transport jobs towards critical operations with no or small float are more urgent and thus will be advantaged over transports to non-critical operations. The combined machine and transport scheduling problem regards constraints (2) and (3) of the JSSP on machine level and constraints (5) and (6) of the *no-wait* JSSP on transport level. In reference to the joint disjunctive graph $G \cup H$ (cf. Figure 3), the joint optimization problem can be stated as follows.

$$\begin{array}{lllr}
\text{Minimize} & TWT + TWC & & (8) \\
\text{s. t.} & S_{io} + p_{io} + t_{io} \leq S_{jp} & O_{io}, O_{jp} \in V,\ \left(O_{io}, O_{jp}\right) \in C & (2) \\
 & S_{io} + p_{io} \leq S_{jp} \mid\mid S_{jp} + p_{jp} \leq S_{io} & O_{io}, O_{jp} \in V,\ \left(O_{io}, O_{jp}\right) \in D & (3) \\
 & S_{kt} + q_{kt} = S_{lu} & T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in E & (5) \\
 & S_{kt} + q_{kt} \leq S_{lu} \mid\mid S_{lu} + q_{lu} \leq S_{lt} & T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in F & (6) \\
 & S_{io} + p_{io} \leq S_{kt} & O_{io}, T_{kt} \in V \cup W,\ (O_{io}, T_{kt}) \in Q & (9) \\
 & S_{kt} + q_{kt} \leq S_{jp} & T_{kt}, O_{jp} \in V \cup W,\ \left(T_{kt}, O_{ip}\right) \in Q & (10)
\end{array}$$

The new objective is to minimize the TWT of the process jobs and the TWC of the transport jobs (8). The raw transport time t_{io} in constraint (2) is ineffective and could be omitted because constraints (5) together with constraints (9) and (10) will ensure that O_{jp} will start no earlier than the end of its last preceding transport operation T_{kt}. The constraints (9) and (10) replace the constraints (7) of the transport scheduling problem. Constraints (9) ensure that the first transport operation T_{kt} of a transport job J_k' can start no earlier than its preceding process operation O_{io} has ended. Correspondingly, constraints (10) ensure that the last transport operation T_{kt} has to be finished before the next process operation O_{jp} can start. Hence, the set Q contains the edges between process operations and first or last transport operations. As a result, set P is not needed anymore for modeling precedences of transport jobs.

The MIP-based scheduling of the transport jobs proceeds as follows, see also Figure 3. First, all machine operations O_{io} (round nodes) that have no unscheduled transport and no unmarked machine operation before them are marked as *done* forming a front line (cross hatched round nodes before f1) and the subsequent transport jobs J_k' are derived. These subsequent transport jobs are checked with each other whether they traverse over the same rotary tables and constitute resource conflicts. Hence, they would have competing transport operations $T_{kt}, T_{lu} \in W,\ (T_{kt}, T_{lu}) \in F$ (square nodes).

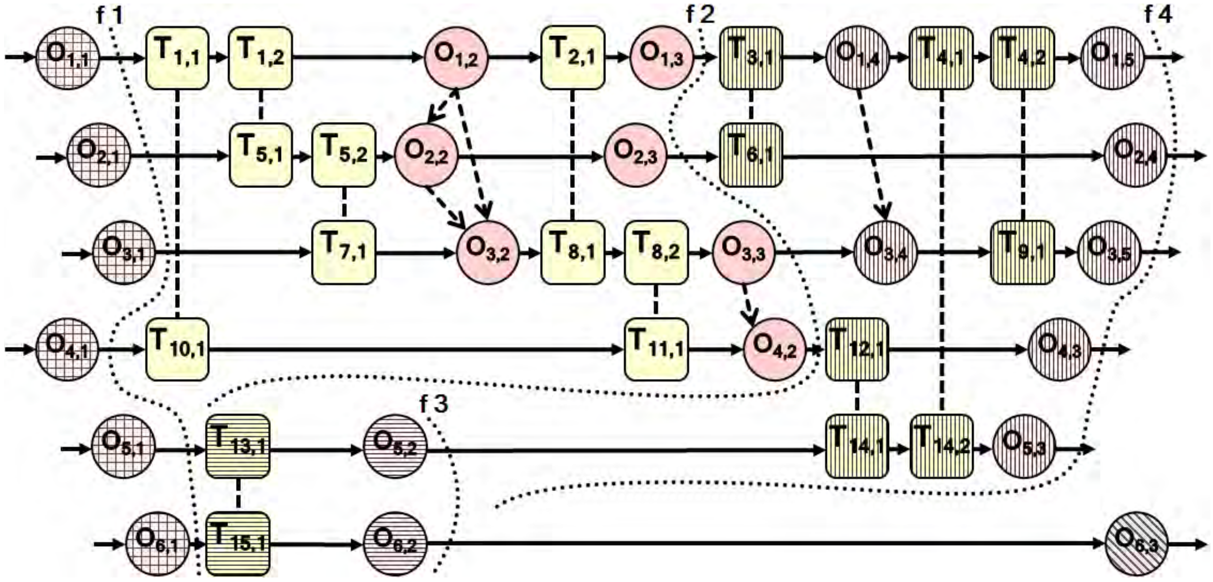

Figure 3: Joint schedule graph $G \cup H$. Sequences (dashed arrows) of machine operations $O_{i,o}$ (circles) are fixed. Competing transport operations $T_{k,t}$ (rectangles, dashed lines) form MIPs to be successively solved.

Due to the fact that some transports might have unhandled machine operations before them, these still marked *undone* machine operations have to be incorporated into the transport scheduling problem as well. In Figure 3, this is the case for the operations $O_{1,2}$, $O_{2,2}$ and $O_{3,2}$. As noted, their sequence (dashed arrow lines) is already fixed, but they might become delayed because of the preceding transport operations conflicts ($T_{1,1}, T_{10,1}$), ($T_{1,2}, T_{5,1}$) and ($T_{5,2}, T_{7,1}$). By checking all currently to be scheduled transports, the transports can be separated into unrelated groups of interrelated transports. For instance, the square nodes between front line $f1$ and front line $f2$ belong to a group of transports that are interrelated due to resource conflicts (dashed lines between plain square nodes). Complementary, the transport operations between front line $f1$ and $f3$ (horizontally hatched square nodes) belong to other transports which are unrelated to the former group of transports. Hence, a transport scheduling problem can be stated separately for each group. Transports of separate groups never visit the same AMHS elements and thus will not constitute resource conflicts. Thus, the size of individual transport scheduling problems is much reduced.

Consecutively, an optimization problem is derived (fully automated) for each of these small groups of interrelated transports. Each of these small optimization problems is solved as a MIP. The MIP solution provides delayed starting times for transport operations S_{kt} as well as for corresponding subsequent machine operations S_{io}. Hence, these delayed starting times must be adopted into the overall (initial) machine schedule and the corresponding target machine operations are labeled *done* (round plain nodes).

To update the remaining subsequent part of the machine schedule, the reduced CPM is used again. Subsequently, the same procedure can be carried out for the next ready transports that are starting from the new front ($f2$) of done machine operations. If the machine schedule is displayed as a graph, a "transport scheduling front" ($f1 \rightarrow f2 \rightarrow f3 \rightarrow f4 \rightarrow \cdots$) rolling forward through the entire graph can be viewed simultaneously with the running computing process of successive MIP solving.

3.4 Machine Schedule Comparison

For each of the three transport scheduling methods one delayed machine schedule has been produced. These three transport-delayed machine schedules are more realistic and thus longer than the undelayed initial machine schedule, see Section 4. More precisely, the initial machine schedule is idealized and could not put into practice because transports in fact do hinder each other in the model as well as in a real AMHS.

4 VALIDATION

The initial machine schedule was constructed for a scheduling horizon of 5 days using product mix and release rates as given in the MIMAC data set. Higher release rates would be possible in principal, but the factory would take in more jobs than its bottleneck machines can process. Thus, higher release rates would lead to eternally increasing fab inventories. Consequently, unrealistic permanent higher release rates were not tested. As a result of the given release rates and planning horizon, 31 jobs with an average of 250 operations using 277 machines, grouped in 97 sets, were produced. A corresponding machine scheduling graph with a total of 7656 nodes was constructed inducing 7625 transport jobs, one after each machine operation except after the final operation of a production job. Each transport consists of a path traversing 10 conveyor elements on average. In general, only 6 of these 10 conveyor elements are rotary tables where wafer lots might compete to enter. In summary, 45672 transport job operations had to be scheduled.

The first delayed machine schedule resulted from the deterministic method *First Come First Serve with Deadlock Avoidance*. As expected, it is the most delayed machine schedule. This is due to the fact that wafer lots wait in queues at the rotary tables. The second delayed machine schedule, obtained by the *Transport Insertion* method, was expected to perform at least as good or better. The third method *Decomposed MIP Solving* was expected to perform best, because it considers several competing transports at once and finds a sequence for these transports that prefers the most urgent transport.

The comparability of the machine schedules is given because of two reasons. First, the initial schedule was constructed regardless of its optimality by a deterministic simulation without any random elements. Second, the initial schedule's structure was never changed, neither during the transport scheduling nor during the feedback of the transport-induced delays. Hence, all compared machine schedules exhibit the same process operation sequences. Due to the fully deterministic nature of the machine scheduling and of all three transport scheduling methods, only one run was carried out for each transport scheduling method.

To ensure the validity, first a sanity check was carried out for each of the obtained schedules. For the machine schedules it was checked, whether each machine is occupied by only a single wafer lot at a time. Furthermore, it was checked whether each process operation actually starts after its preceding operations. In the same manner the transports jobs were checked. First, it was checked whether the transport operations are carried out in the correct order and without delays. Furthermore, it was checked, if each transport starts after its preceding source machine operation has finished and whether its succeeding target machine operation starts after the transport has arrived.

For assessment of the delayed machine schedules, first cycle times as the time between job release and job completion had to be computed and averaged. Second, the tardiness as the difference between due date and completion time was computed. The due date factor for this assessment was 1.0. Hence, this tardiness exhibits exactly how tardy a process job became due to the transport-induced delays. Third, the percentage improvement was calculated, whereas the state-of-the-art method *First Come First Served with Deadlock Avoidance* (FCFS DA) served as the baseline. More precisely, the saved delays were calculated as the difference of baseline total tardiness (2339) minus the reduced AMHS-induced tardiness. Then this value was put in relation to the baseline tardiness.

Table 1: Improvement comparison of transport scheduling methods.

Transport Scheduling Method	Average Job Cycle Time in dd, hh:mm:ss	Total AMHS-induced Tardiness in Seconds	Improvement vs. Baseline in %
Method 1 (FCFS DA), Baseline	10, 16:34:30	2339	0%
Method 2 (Transport Insertion)	10, 16:33:42	835	64.30%
Method 3 (MIP-Based Solving)	10, 16:33:28	420	82.04%
None, no AMHS (unrealistic)	10, 16:33:15	0	100.00%

The total amount of the AMHS-induced tardiness can be judged as marginal. This small amount of AMHS-induced tardiness is due to two reasons. First, cases of impeding are curtailed due to the AMHS model's double-tracked unidirectional conveyors. Second, the traffic loads were comparatively moderate. But the improvement percentages clearly show the potential of continuous transport flow scheduling.

5 CONCLUSION

In this study, three different transport scheduling methods have been investigated in order to identify which one works best in principle for a proof of concept. A transport scheduling method is considered to work well if it avoids transport-related delays as much as possible. Transport-related delays are caused by one wafer lot impeding another because they happen to appear at the same rotary tables at intersecting conveyors. A realistic machine schedule that considers these transport-related delays will be delayed compared to an ideal machine schedule. An ideal machine schedule does not consider transport delays, but assumes ideally short raw transport times instead. In contrast, we suggested and tested an approach that feeds the transport-related delays back into an initially given ideal machine schedule. As a result a more realistic (although delayed) machine schedule is obtained.

The first transport scheduling method resembles the FCFS-oriented control logic as it is installed in most real state-of-the-art wafer fabs with a CFT-based AMHS. Therefore, this method served as the baseline for two more advanced methods. The former of these two advanced methods constructs a transport schedule by successively inserting complete transports so that the first available suitable time window will be used. The latter method first identifies groups of transports that may interfere with each other and then resolves these conflicts by deriving a small MIP and solving it optimally. These MIPs consider the critical operations of the superordinate initial machine schedule for preferring the corresponding critical transports. Hence, these MIPs exploit CPM-computed floats for disadvantaging transports before non-critical machine operations. As expected this MIP-based method produces the best delayed machine schedules.

In the dynamic scheduling context of real fabs, new jobs constantly enter. Hence, schedulers prefer fast scheduling methods that use a set of dispatching rules for constructing the machine schedule. Often this is done synchronized with the real production and the decisions about which wafer lot to assign to an idle machine are done not much earlier than the real machine actually becomes idle in reality. In such a setting the *Transport Insertion* method can be applied because it only considers the transport at hand when a machine operation is finished and then it determines a starting delay for this transport. For determining this delay, a schedule of all previously started transports has to be maintained which can be done in a central database.

If the *MIP-based* transport scheduling method shall be applied, a transport job forecast is needed. This transport job forecast could be achieved by ongoing in-process simulation of the machine dispatching rules. In contrast to simulation, advanced heuristics (Mason et al. 2002) or even exact methods (Klemmt et al. 2009) are arising for use in practice. In these cases machine schedules would be known for a given planning horizon. Thus, *MIP-based* transport scheduling method can be applied. The only prerequisite is that the average transport duration is significantly smaller than the planning horizon. Otherwise, the new planning period could introduce new transports that interfere with transports that just have been scheduled and fixed for the old planning period. Special attention has to be paid if rescheduling of machine operations is done. The corresponding possibly already scheduled transports then have to be rescheduled as well. Transports that already physically started, but might interfere with newer not yet started but to be rescheduled transports, have to be considered fixed. Precisely, the MIP solver has to (re)schedule the newer transports “around” the older fixed transports.

In summary, the *Transport Insertion* method could be integrated easier into control systems of existing wafer fabs, but the *MIP-based* transport scheduling method is more beneficial and should be pursued along with advanced fab scheduling methods. Accordingly, future work will apply the methods to more optimal, i. e. denser, fab schedules, which induce higher traffic loads. Thus, more simulation runs using other fab models and covering longer periods of time shall show the benefits of CFT scheduling more clearly.

ACKNOWLEDGMENTS

This work was partly funded through the research projects SemI40 under agreement # 16ESE0077S and PowerBase under agreement # 662133 both funded by ECSEL Joint Undertaking (European Union's Horizon 2020, BMBF, SMWA), RaProUm funded by BMWi under agreement # ZF4098802LP6 and TOPAs funded by the research program European Union’s Horizon 2020 under agreement # 676760.

REFERENCES

Agrawal, G. K., and S. S. Heragu. 2006. “A Survey of Automated Material Handling Systems in 300-mm Semiconductor Fabs”. In *IEEE Transactions on Semiconductor Manufacturing*, 19 (1), 112–120.

Arzt, T., F. Bulcke. 1999. “A New Low Cost Approach in 200 mm and 300 mm AMHS”. In Semiconductor Fabtech 10, 19–26.

Bannert, A., F. Heinlein, M. Adam, and K. Manja. 2012. “Operator-free Exception Measurement Logistics for a Highly Automated 200mm Semiconductor Manufacturing Environment”. In *Advanced Semiconductor Manufacturing Conference (ASMC) 2012, 23rd Annual SEMI*, 251–256.

Brain, M., R. Gould, U. Kaempf, and B. Wehrung. 1999. “Emerging Needs for Continuous Flow FOUP Transport”, In *Electronics Manufacturing Technology Symposium 1999, 24th IEEE/CPMT*, 76–82.

Deroussi, L., M. Gourgand, and N. Tchernev. 2008. “A Simple Metaheuristic Approach to the Simultaneous Scheduling of Machines and Automated Guided Vehicles”. In *International Journal of Production Research* 46 (8), 2143–2164.

Drießel, R., L. and Mönch. 2012. “An Integrated Scheduling and Material-Handling Approach for Complex Job Shops: a Computational Study”. In *International Journal of Production Research* 50 (20), 5966–5985.

Fowler, J. W., and J. Robinson. 1995. “Measurement and Improvement of Manufacturing Capacities (MIMAC): Final Report”. Technical Report No. 95062861A-TR, SEMATECH, Austin, TX.

Graham, R. L., E. L. Lawler, J. K. Lenstra, and A. R. Kan. 1979. „Optimization and Approximation in Deterministic Sequencing and Scheduling: A Survey”. In *Annals of Discrete Mathematics* 5, 287–326.

Hammel, C., T. Schmidt, and M. Schöps. 2012. “Network Optimization Prior to Dynamic Simulation of AMHS”. In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach,. R. Pasupathy, O. Rose, and A. Uhrmacher, 172. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Heinrich, H., and A. Pyke. 1999. “The Impact of Conveyor Transports on Factory Performance at Infineon’s (Siemens) 200 mm fab”. In Semiconductor Fabtech 10, 15–18.

Heinrich, H., G. Schneider, F. Heinlein, S. Keil, A. Deutschländer, and R. Lasch. 2008. “Pursuing the Increase of Factory Automation in 200mm Frontend Manufacturing to Manage the Changes Imposed by the Transition from High-Volume Low-Mix to High-Mix Low-Volume Production”. In *Advanced Semiconductor Manufacturing Conference, ASMC 2008. IEEE/SEMI*, 148–155.

Hong, S., A. L. Johnson, H. J. Carlo, D. Nazzal, and J. A. Jimenez. 2011. “Optimising the Location of Crossovers in Conveyor-Based Automated Material Handling Systems in Semiconductor Wafer Fabs”. In *International Journal of Production Research* 49(20), 6199–6226.

Jimenez, J., G. Mackulak, and J. Fowler. 2008. “Levels of Capacity and Material Handling System Modeling for Factory Integration Decision Making in Semiconductor Wafer Fabs”. In *IEEE Transactions on Semiconductor Manufacturing*, 21 (4), 600–613.

Jiong, Z., W. Yu-bao, Z. Jie, and L. Si-jiang. 2013. “On the Active Control Properties of Branching Nodes in Complex Conveyor Systems”. In *Proceedings of the 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, 3257–3261. Piscataway, NJ: IEEE, Inc.

Klemmt, A., G. Weigert, C. Almeder, and L. Mönch. 2009. “A Comparison of MIP-based Decomposition Techniques and VNS Approaches for Batch Scheduling Problems”. In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1686–1694. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Lacomme, P., M. Larabi, and N. Tchernev. 2013. “Job-Shop Based Framework for Simultaneous Scheduling of Machines and Automated Guided Vehicles”. In *International Journal of Production Economics* 143 (1), 24–34.

Lasrado, V., and D. Nazzal. 2011. “Design of a Manufacturing Facility Layout with a Closed Loop Conveyor with Shortcuts using Queueing Theory and Genetic Algorithms”. In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach,. K. P. White, and M. Fu, 1964–1975. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Mason, S. J., J. W. Fowler, and W. M. Carlyle. 2002. “A Modified Shifting Bottleneck Heuristic for Minimizing the Total Weighted Tardiness in a Semiconductor Wafer Fab”. In *Journal of Scheduling*, 5 (3), 247–262.

Mönch, L., J. W. Fowler, S. Dauzere-Peres, S. J. Mason, and O. Rose. 2011. “A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations”. In *Journal of Scheduling* 14 (6), 583–599.

Mönch, L., J. Fowler, S. J. Mason. 2012. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. New York: Springer.

Montoya-Torres, J. 2006. “A Literature Survey on the Design Approaches and Operational Issues of Automated Wafer-Transport Systems for Wafer Fabs”. In *Production Planning and Control* 17 (7), 648–663.

Nazzal, D., J. Jimenez, H. Carlo, A. Johnson, and V. Lasrado. 2010. “An Analytical Model for Conveyor-Based Material Handling System with Crossovers in Semiconductor Wafer Fabs”. In *IEEE Transactions on Semiconductor Manufacturing*, 23 (3), 468–476.

Paprotny, I., J.-Y Shiau, Y. Huh, and G. Mackulak. 2000. "Simulation Based Comparison of Semiconductor AMHS Alternatives: Continuous Flow vs. Overhead Monorail". *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 1333–1338. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Pettinato, J. S. and D. Pillai. 2005. "Technology Decisions to Minimize 450-mm Wafer Size Transition Risk". In *IEEE Transactions on Semiconductor Manufacturing* 18 (4), 501–509.

Poppenborg, J., S. Knust, and J. Hertzberg. 2012. "Online Scheduling of Flexible Job-Shops with Blocking and Transportation". In *European Journal of Industrial Engineering* 6 (4), 497–518.

Qu, P., B. Steinmiller, and S. J. Mason. 2003. "Incorporating Automated Material Handling Systems into a Disjunctive Graph for Subsequent Scheduling by a Shifting Bottleneck-based Approach". In *IIE Annual Conference. Proceedings*, Institute of Industrial and Systems Engineers (IISE), 1-5.

Scholl, W., B.-P. Gan, P. Lendermann, D. Noack, O. Rose, P. Preuss, and F. Pappert. 2011. "Implementation of a Simulation-based Short-term Lot Arrival Forecast in a Mature 200mm Semiconductor Fab". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach,. K. P. White, and M. Fu, 1927–1938. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Sourirajan, K., and R. Uzsoy. 2007. "Hybrid Decomposition Heuristics for Solving Large-Scale Scheduling Problems in Semiconductor Wafer Fabrication". In *Journal of Scheduling* 10 (1), 41–65.

Temponi, C., J. A. Jimenez, and F. A. M. Mediavilla. 2012. "Critical Variables in the Decision-making Process for AMHS Technology Selection in Semiconductor Wafer Size Transitions: Exploratory Study". In *IEEE Transactions on Semiconductor Manufacturing* 25 (3), 408–419.

Wang, C.-N., Y.-T. Chung, Y.-H. Wang, M.-T. Duong, and T.-F. Lin. 2016. "The Material Dispatching Method for Conveyor System in 450 mm Wafer Fabrication". In *Journal of Testing and Evaluation* 45 (3).

Wang, I.-L., Y.-C. Wang, and C.-W. Chen. 2013. "Scheduling Unrelated Parallel Machines in Semiconductor Manufacturing by Problem Reduction and Local Search Heuristics". In *Flexible Services and Manufacturing Journal* 25 (3), 343–366.

Zhang, J., W. Qin, and L. Wu. 2016. "A Performance Analytical Model of Automated Material Handling System for Semiconductor Wafer Fabrication System". In *International Journal of Production Research* 54 (6), 1650–1669.

Zhang, Q., H. Manier, and M.-A. Manier. 2012. "A Genetic Algorithm with Tabu Search Procedure for Flexible Job Shop scheduling with Transportation Constraints and Bounded Processing Times". In *Computers and Operations Research* 39 (7), 1713–1723.

Zhou, B.-H., J.-X. Chen, and Z.-Q. Lu. 2016. "An Analytical Model for Continuous Flow Transporters of AMHSs with Multi-loop Conveyors and Priority Rules". In *International Journal of Computer Integrated Manufacturing* 29 (5), 489–503.

AUTHOR BIOGRAPHIES

CLEMENS SCHWENKE received his M. S. degree in Electrical Engineering from Dresden University of Technology and he is now a PhD student at the Chair of Technical Information Systems of Professor Klaus Kabitzsch. His research interests include modeling, simulation and scheduling in automation. His e-mail is clemens.schwenke@tu-dresden.de.

KLAUS KABITZSCH holds the Chair of Technical Information Systems at the Institute of Applied Computer Science of the Dresden University of Technology, Germany. He received a Diploma and a PhD in Electrical Engineering and Communications Technology. His current projects focus on software tools for design of networked automation, data analysis, advanced process control and predictive technologies. He is a member of IEEE, VDE and GI. His e-mail is klaus.kabitzsch@tu-dresden.de.