Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

STATIC AMHS SIMULATION BASED ON PLANNED PRODUCT MIXES

Robert Schmaler
Christian Hammel

FabFlow GmbH
Pohrsdorfer Weg 3A
01169 Dresden, GERMANY

Christian Schubert

Infineon Technologies Dresden GmbH & Co KG
Königsbrücker Str. 180
01099 Dresden, GERMANY

ABSTRACT

Automated material handling systems (AMHS) in semiconductor fabrication plants (Fabs) are crucial to achieving a high production throughput. When upgrading an existing or planning a new Fab, production plans are used to decide on the required toolset. But what about AMHS planning? To our knowledge no method exists, given an anticipated product mix, that generates reliable transport patterns including non-productive transports. This paper outlines a methodology to generate such transport patterns, including non-productive transports for test wafers, empty FOUPs, and others. The prediction is based on a preliminary product mix supplemented with scheduling rules and AMHS characteristics, yielding deeper insight into actual AMHS capabilities and constraints during the planning phase. The results are used as input to a Static Simulation which will create a forecast on track utilization for a given layout in which possible AMHS bottlenecks are highlighted.

1 INTRODUCTION

A well designed AMHS (Automated Material Handling System) moves carriers seamlessly between process steps while not limiting tool utilization, i.e. there are no transport induced idle times. Thus, the AMHS has to meet tool-specific Carrier Exchange Times (CET) (Rothe et al. 2015) for all tools and over the whole lifetime of the Fab. Future Fab scenarios are planned based on an anticipated product mix. To incorporate possible changes to the AMHS layout, it is crucial to know how track-utilization is going to change as early as possible. However, so far future transport patterns have mostly been generated based on experience and (vague) assumptions only.

We are considering modern OHTs (Overhead Hoist Transport) representing unified systems that are built in a bay-chase layout (Murray et al. 2000). These unified AMHS enable inter- and intra-bay-transports of FOUPs (Front Opened Unified Pods) without switching between vehicles. Since this allows for direct tool-to-tool transports between different modules, transport patterns become more important. One has to be aware of where to place high throughput tools, have the additional traffic in this area in mind and compare this to AMHS track capacities, especially at critical points like intersections. During the last years, stockers as mass storages are used less frequently but can still be found in most Fabs. They are often placed inside the Interbay for easier access and less interference with process tools. Since Interbays are usually the most utilized parts of the OHT network, thorough analyses should be conducted to avoid unexpected bottlenecks.

Newly built Fabs still have numerous degrees of freedom regarding track layout and tool placement. It becomes more involved for existing Fabs with plans to upgrade their AMHS and options are very limited for Fabs in full production. Adapting to significant changes in production scenarios could, therefore, be limited by existing AMHS bottlenecks or even create new ones. The earlier these restrictions are known, the earlier possible countermeasures can be designed and implemented. Thus, time plays a crucial role in designing and adapting the AMHS as well as influencing the position of high throughput devices or even relocate existing process tools.

To generate transport data, detailed information about the product mix is necessary. Fab simulation can then try to calculate the tool utilization and decide about the toolset. A wafer has to pass hundreds of process steps and, at least in highly utilized Fabs, only a fraction of all transports are direct tool to tool transports. To extract reliable information about future track utilization, it is crucial to predict where lots will be transported to and stored in between process steps. Moreover, additional transports, e.g. test wafers, reticles, or NPW (Non-Productive Wafer) material have to be taken into account.

The presented approach to calculate a From-To-Matrix of future transports was developed for an existing Fab with potential changes in product mix and layout. It is based on generating transport patterns directly within RTD (Real-Time Dispatcher by Applied Materials) which allows access to a broad spectrum of Fab data. It has been built in two iterations to ensure validity: First, a model was developed of the current layout, tested, and validated against the running system. Second, this validated model was used to forecast transport patterns of future product mixes.

Figure 1 illustrates the general idea behind this approach. The necessary input data consists of the already mentioned product mix, tool configuration information (including recipe depending process times and process definitions to define possible tools for each process step), and experience of the developer. This is important since not all transports can be generated based on the production plans only. This input results in a From-To-Matrix which subsequently will be the input to a simulation. Together with the future layout, a Static Simulation can be conducted yielding a track utilization chart to identify potential bottlenecks. Another possibility is to perform a Dynamic Simulation, but this topic will be discussed in a follow-up paper.

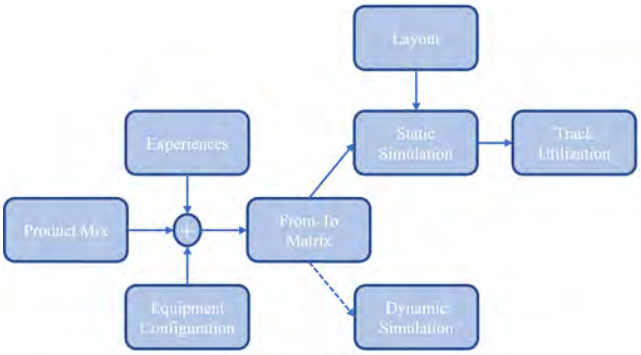


Figure 1: Generation of transport matrix and further AMHS analysis

In most cases, a Dynamic Simulation of the AMHS is conducted to answer the relevant questions. Since this approach is very time consuming this paper uses an optimization approach before Dynamic Simulation. As already introduced in Hammel et al. (2012), the Static Simulation approach uses the representation of the AMHS as a network-model. Already existing sophisticated algorithms can be adopted. The improvements possible with this approach applied on a running Fab with historical transport data can be found in Hammel et al. (2012) with significant results when validating the results of the Static Simulation with a dynamic one (20% reduction of delivery times, 20% increase of max throughput without a considerable increase in delivery times).

The rest of this paper is organized as follows: first, a summary of known research will be given which deals with transport job generation out of product mixes. Afterward, a methodology will be presented defining the necessary input data and how transport forecasts for both, existing and future Fabs, can be derived from it. The results will show details about the validation process based on real-life data. The resulting transports are transferred into the layout using FlowLogiX MMS software. Additional results of optimization of cost settings and referring routes are also presented. This whole process is shown in Figure 2. Finally, the paper will be summarized together with an outlook on future tasks and room for improvements.

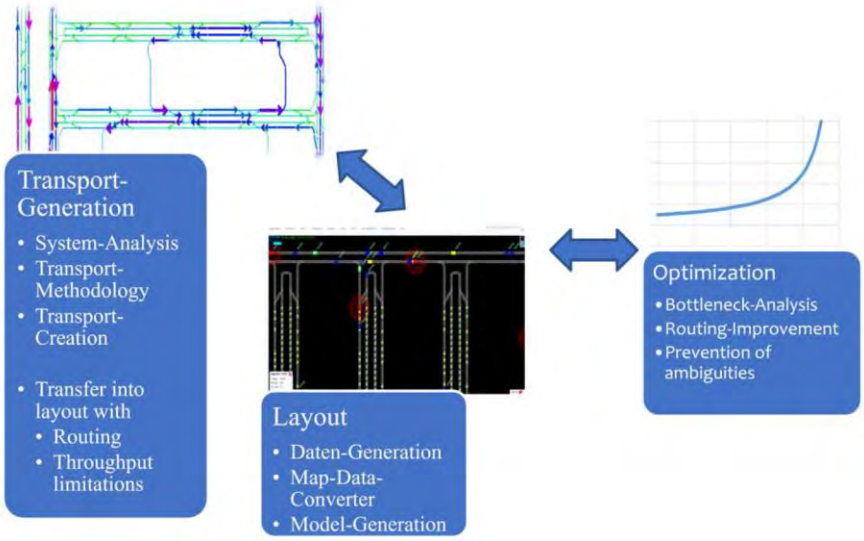


Figure 2: From transport generation to AMHS optimization with FlowLogiX MMS (Monitoring and Maintenance Suite)

2 LITERATURE REVIEW

To gain insight into track utilization and capacity of the OHT most literature sources use transports observed from running systems when there is already an existing Fab to deliver data. Only a few publications are dealing with other approaches which are necessary when the Fab is still to be built. In Gaxiola et al. (2013) a linked AMHS and Fab model is used with the focus on the Fab model incorporating travel time statistics from the AMHS model. To generate transports in the AMHS model, the authors used planning information (i.e. process flow, start rates, and tool lists) as the predictions lack real transport data. All NPW movement is taken into account by upscaling the productive transports by a safety factor of 1.8x. Since there is no existing Fab to compare to, no information about the accuracy of this transport generation has been provided. In Lin, Chao-Jung (2014) the AMHS of a photo bay in a 300mm wafer Fab is modeled to find optimal dispatching rules and the number of vehicles. Transport demands are modeled by calculating inter-arrival rates of wafers regarding WSPW (Wafer Starts per Week), the process flow, and the number of processes entering the photo bay. This is similar to the assumption of a steady-state scenario as described later but focusing on production material only. In Kiba et al. (2009) simulation models of the production system and AMHS are combined to a full Fab model. They include lot scheduling as well as lot storage to model the production system. Nevertheless, they neither handle the questions of NPW nor of the accuracy of their transport job generation approach. They also focus on the problem of storage capacity which is also addressed in Dauzere-Péres et al. (2012). The latter aims at finding the right default storages with the capacity depending on tool grouping (based on the same process steps) and tool throughput. A MILP (Mixed Integer Linear Programming) approach minimizing travel distances between tools and storages as well as minimizing maximum storage utilization is presented. Again, NPW transports have been explicitly neglected. Additionally, their MILP to group and set capacities of storages is not able to find optimal solutions in a reasonable amount of time for a large-scale wafer Fab as regarded in this paper, unfortunately.

When it comes to use the generated transports and check AMHS capability, most researchers make use of Dynamic Simulation. Nevertheless, Dynamic Simulation needs adequate input data to gain from its benefits (e.g. vehicle interaction). Static Simulation uses the fact that routing in AMHS is mostly static (see Lee 2016). That means that vehicles always use the same route between any given source and destination due to static cost settings. Bartlett et al. (2014) differentiate between static and dynamic routing approaches, too. But they define static routing as the selection of the route at the beginning of a transport meaning no

update or re-calculation during transport execution. Dynamic routing on the other hand continuously re-evaluates the current route of the OHT vehicle. The presented static approach uses cost settings based on minimum or average historical transport times. They adjust the cost settings due to current transport times in repeating intervals (dynamic pre-planned routing). It is no systematic optimization of the setting regarding the whole transportation system, it's more likely the common approach of iteratively increasing costs of over-utilized links. Lee (2018) mentions the possibilities of re-routes and therefore the relaxation of over-utilized pieces of track. In the first paper, they are talking about adjusting the cost settings by Fab managers in a static way. Later they are using traffic jam recognition values as relative congestion index (RCI) to find tracks where cost settings have to be adjusted and adjust the values based on their calculations. But they do not answer the question about the impact of changing one value on the rest of the system.

The only known approach about static (pre-)optimization of cost settings of the whole system is the one from Hammel et al. (2012) which will be summarized in chapter 4.

All in all, there is a lack of validation of calculated transports compared to real transports since most Fabs do not yet exist. The main differences of the approach presented here are the generation of the transport lists directly in RTD, without any other additional simulation software like AutoSchedAP, and always using the same rules as in the "real world" and the fact that NPW transports are taken into account. Additionally, it is possible to use the calculated results together with the identified rules of transport job generation to evaluate real Fab performance (e.g. regarding storage allocation). Finally, the use of Static Simulation of the transport system results in fast and reliable insights in future track utilization and AMHS bottlenecks with the possibility of overall cost optimization.

3 METHODOLOGY OF FROM-TO GENERATION

This section summarizes the generation of the From-To-Matrix. For a more detailed description, please refer to Schmaler et al. 2020.

Within the scope of this paper, the required transport data is summarized by a From-To-Matrix $A = \{a_{ij}\}$. Each row of this matrix describes a potential source i, while each column represents a potential destination j of a transport. The elements a_{ij} give the number of transports going from i to j within a given time frame, e.g. a week.

Since a steady-state approach is used, no information regarding time-correlation between these transports has been calculated. Although for batch tools, for example, the average transport load might be small, the actual load may display sharp spikes over time since many transports could be carried out simultaneously with long pauses in between. We assume that this effect will be less significant on the track load, as there are many tools connected to it which will lead to a more evenly distribution of transport load. Nevertheless, these dynamic effects have to be accumulated into a safety factor *S,* which needs to be applied when comparing the required transport volume with given AMHS capabilities. Moreover, the steady-state assumption also ignores potential uneven WIP distribution, potentially caused by planned or unexpected tool downs. Since the planned product mix normally, assumes a high tool utilization there is only a limited capacity for tools to manufacture more lots than planned, thus the effects on transport load may be considered limited. Still, the aforementioned safety factor S has to account for these effects since they might not be insignificant.

To calculate the complete From-To-Matrix *A*, we assume that we have access to the future product mix. The product mix shall be given by some process flows pf_k along with their wafer start rate ws_k. Each process flow pf_k itself contains a sequence of process steps $p_{k,l}$ that need to be executed, where each process step can be carried out by a given set of tools $ts_{k,l}$. Please note that the set of tools $ts_{k,l}$ qualified to perform a process step might be considered a variable itself. For our calculations, we shall consider it to be given as this is a common problem and has been addressed elsewhere. It is worth noting, that certain process steps may be sampled, i.e. only a given percentage of lots traveling on this process flow will perform this step. This is being described by a sampling rate $sr_{k,l}$. A common example is measuring steps to identify potential defects as early as possible. Another challenge is to consider conditional process steps, i.e. steps

that are only executed under certain circumstances. To arrive at reliable transport data, good estimates have to be found for that. Ideally by looking into data from similar existing Fabs.

During the planning phase, often only the product mix for the products that shall be manufactured in the future is known. However, transports caused by productive wafers only account for a fraction of total transports. Therefore each transport type will be looked at in the following subsections.

3.1 Production wafers

Given the product mix, one can easily obtain the number of carriers $c(p_{k,l})$ per time unit that is being processed at each process step given the average number of wafers per carrier w_k on this process flow.

$$c(p_{k,l}) = ws_k \frac{1}{w_k} sr_{k,l}$$

Since the product mix uniquely defines the previous as well as the next process step and their respective toolsets (except for sampling and conditional steps) carrier transports can be derived by choosing appropriate tools from this toolset. Please note, that this is a non-trivial task because tools may be part of multiple toolsets and thus may occur multiple times on the same or different process flows. Also, tools may exhibit different process times within a given toolset which is why this should either be solved by an optimization algorithm or at least by an intelligent dispatching algorithm that properly balances tool load.

Having established how many carriers travel between which tools, the next step is to augment these transports as only a few will travel directly from tool to tool. In most cases, they will be stored temporarily at least once on their way, depending on the storage policies in place, see Hammel et al. (2016). Another example is tools with a very short CET (carrier exchange time) which might have designated fast access storages near the tool, causing additional transports.

When choosing storages, similar balancing rules, e.g. concerning their turnover frequency, as in the case of choosing tools from a toolset have to be applied since real-time information on available capacity is not available in steady-state calculations.

3.2 Reticle transports

Since reticles are an auxiliary resource required to process wafers during lithography processes, their transport numbers can also be derived from the product mix. They are either send from a storage location to a tool, from a tool to another tool or back to storage. Ideally, lots are dispatched to litho tools such that reticles may be used for many lots thereby reducing the overall number of required reticle transports. The effectiveness is measured by the reticle factor F_{Ret} defining the average number of lots being processed per reticle:

$$F_{\mathrm{Ret}} = \frac{Lots}{\mathrm{Re}\,ticle}$$

Using this ratio, the corresponding reticle transports can be generated by looking at the assigned carrier transports for each tool.

3.3 Empty FOUP transports

Empty FOUPs are necessary at various process steps. They are required for the creation of split lots or to avoid contamination, for example. Since the data has been created directly from RTD, runtime configuration of the usage of empty FOUPs has been available which could be extrapolated to future processes by defining template tools and processes.

3.4 Test wafer

Test wafers are used to continuously provide feedback on process stability in different ways. Some regularly assess certain tool characteristics and therefore scale with the number of processing tools. Others frequently

gather in-depth data on specific processes and therefore scale with the number of processes per tool, the number of tools, and their throughput. If we assume that all of these quantities display a linear relationship with respect to the combined wafer starts $ws = \sum ws_k$ then, by applying superposition, we can assume that the total number of test wafers is also linear with respect to ws.

Since many test wafers need to be conditioned or even manufactured before the actual test can be performed, it is reasonable to assume that they follow certain test wafer specific process flows, as well. Obtaining transports can thus be done, similar to production wafers, by either assigning existing or designing new process flows.

4 METHODOLOGY OF STATIC SIMULATION

The previous chapter described the methodology of a simplified Fab simulation to generate a from-to matrix out of a product mix and necessary (AMHS-) routing rules. The goal was to achieve reliable information about AMHS load due to the used product mix. Next, this transport information has to be transferred into a layout model to get the resulting track utilization. A promising approach for a simulation at an early stage is a static one.

In a Static Simulation, an AMHS is represented by a network graph as described in Hammel et al. (2012). This is possible since the routing in complex AMHS is mainly static as availability of information is insufficient for a dynamic approach. Most systems are based on a shortest-distance path routing. That results in each transport from source to destination taking the same path, regardless of the current traffic (exceptions are special traffic dependent functions as traffic jam avoidance that increase cost settings dynamically).

Static Simulation uses methods from Graph Theory. That means that the AMHS track, the storages, and the tool ports are translated into links (representing the track) with the attached source and sink information (representing the load-ports). On the one hand, the advantage is that the shortest paths are easy to find and sophisticated algorithms already exist to identify them. On the other hand, a disadvantage is, that no dynamic behavior can be represented by this. As a result of the network approach, track and station utilization of the forecasted transports can be easily shown within the layout.

Track utilization uses the basics of betweenness centrality from the Theory of Complex Networks (Newman 2003, Bocaleti et al. 2006). The idea is to model the average transports per unit of time (e.g. per hour) and therefore transports as flows. To cope with the missing dynamic behavior, the utilization uses a lower limit than the technical limit of the system to represent a buffer for dynamic behavior. If all tracks keep this limit, congestions due to traffic should be rare and the impact of failures should be lower, resulting in higher robustness of the system. In running systems or planned ones where the layout is already fixed, virtually adjusting the lengths (costs) of links enables manipulating routing with no or minor software or hardware changes. An analytic approach to generate link costs in a way that keeps all throughput limits is not feasible because of run time. That is why an iterative algorithm is used to increase the costs of over-utilized links. The limits of the different parts of the systems might be defined by the vendor or user experience. Different limits might be necessary for different links (e.g. OHT track vs. conveyor parts). The iterative approach increases cost settings one by one or all over-utilized links at once. The amount to increase depends on the over-utilization. The amount of utilization lowering and the mean shortest path length increase also influence the iterations. Figure 3 shows an easy example of how adjusting the cost settings can change transport routes.

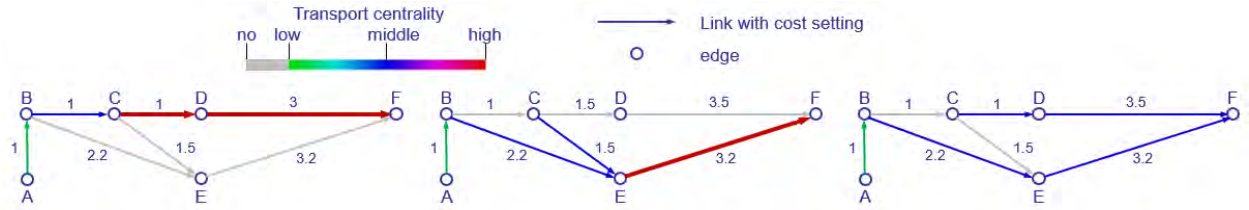


Figure 3: example network showing transport centrality with initial settings (left), with changing multiple link costs (middle) and with changing the cost of a single link (right)

Compared to the more common Dynamic Simulation approaches the Static Simulation has multiple advantages for the use case described in this paper:

- Reduced effort for model building and adjustment
- Efficient analysis and summary of routes and expected track usage
- Simple alternative testing
- Intuitive macroscopic view onto expected system behavior
- General possibilities of analyzing relevant relations and inducted impact on potential bottlenecks
- Possibility to focus on relevant areas within Dynamic Simulation

For the applicatory part of the Static Simulation, a software package called FlowLogiX MMS is used. The FlowLogiX Maintenance and Monitoring Suite is a software for monitoring and analyzing AMHS, to build a Static Simulation resulting in track utilization and identifying current or future bottlenecks. Besides the optimization of the cost settings and the visualization, the import of the original layout is the main challenge. Figure 4 shows the individual steps of this process. At first, all necessary data has to be already entered into an origin CAD layout. The next step is to extract the important information about segments, load ports, and buffer from this CAD file. As a result, for each of the extracted system parts, different information is available: name, type, coordinates, and attributes (e.g. direction). An additional adapter converts the information into an FLX map data file which translates it all into one network model.

The model can also be enriched with track capacities (based on various types of transport systems like OHT, conveyor, lift, and including buffer for variation over time and empty vehicle movements). Also, the software allows static routing optimization for better utilization of the transport network. For layout planning application this can result in four different scenarios: (1) the layout can handle transport requirements, (2) routing optimization can ease bottlenecks, (3) adjustments in track layout have to be made or (4) adjustment in tool layout are necessary.

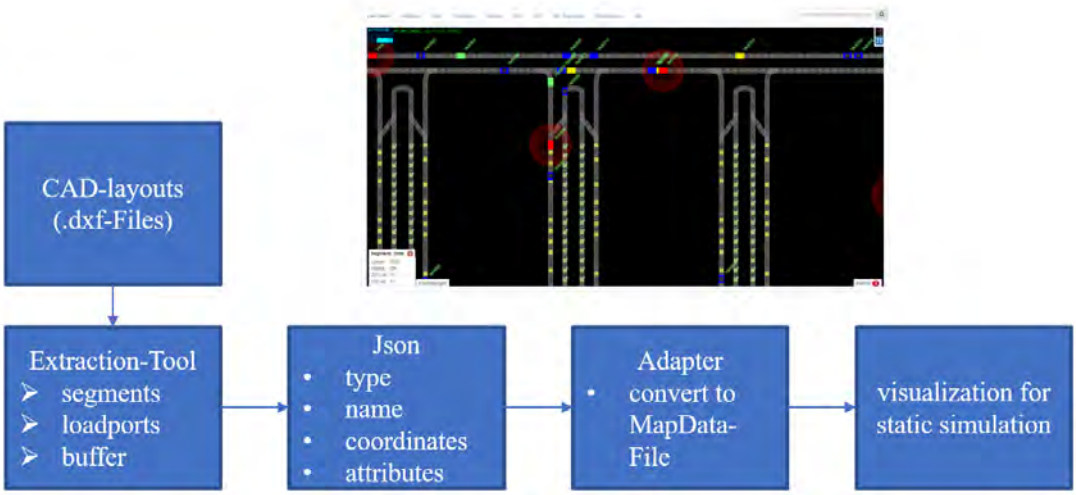


Figure 4: Layout-import with FlowLogiX MMS

5 RESULTS

The following results will discuss the transport generation first and the Static Simulation of the transports with optimized cost settings afterward. The validation process during the From-To-Matrix calculation showed that it is possible to generate a reliable matrix using the described approach. Different settings for different inputs have been tested during the development of the from-to matrix generation process. The first part was to find accurate product mix input data. The problem during validation using historical product mix is that this already generates an initialization error since the real wafer starts may differ from the forecasted ones. To overcome this problem the real wafer starts based on the wafer start plan of the analyzed time period have been used. So the total amount of process steps was guaranteed to be the same in reality as in the transport generation. This has also been recommended by the internal Fab simulation team.

The selection of specific tools was based on item time (recipe dependent average time to handle one wafer) with the goal of even tool utilization. An accuracy of choosing a tool in the correct BCU (Bay Control Unit) of above 90% could be achieved. The main deviation in tool selection was observed for measuring tools. This has two reasons. Firstly, these are the most distributed ones and secondly, the groups to choose the right tool for the next process step are the largest.

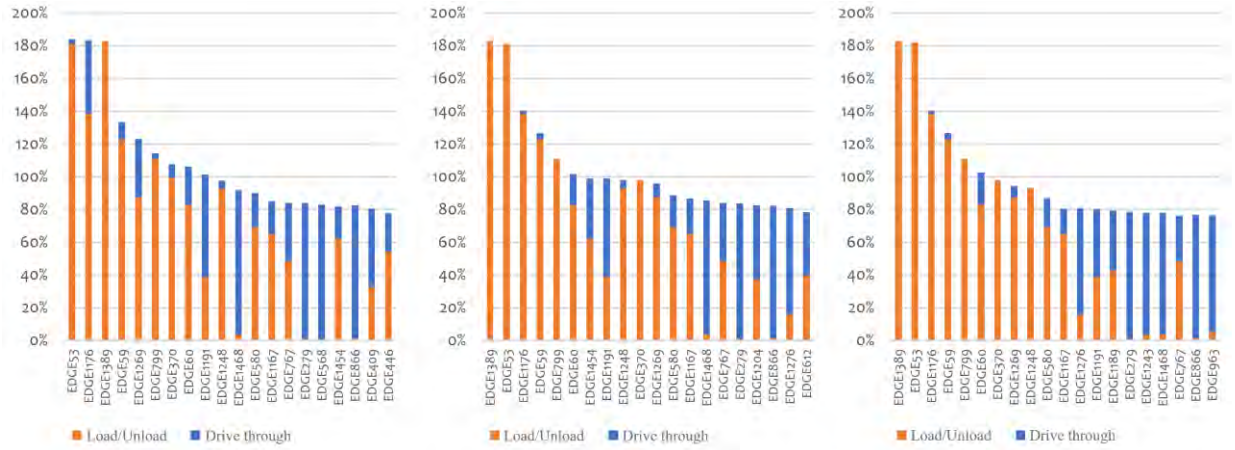
To identify transport relations with the highest deviations from real transports the calculated From-To-Matrix was directly compared to real Fab transports. In the end, the total transport volume has been matched with an accuracy of 96%. The From-To-Matrix showed an accuracy of more than 80% matching source-to-sink relations of real transports (productive and non-productive). As the approach is a static one to allow fast analysis and not considering current WIP distribution or available storage capacities these values show very good accuracy. The results are even more promising since the tools of the same group are not all situated in the same bay. Instead, there are some which are quite distributed over different bays.
The reasons for deviations in the control matrix are:

- **Wrong storage selection** – The implemented where-next rule already may itself not always pick the right storage (e.g. in case of unexpected tool downs). By definition, the forecast cannot exceed the accuracy of the where-next rule from reality. Luckily some “wrong” storages are still located in the right BCU which reduces the impact on accuracy on BCU level. The second reason for the wrong storage selection is that it is hard to calculate a total storage capacity over a fixed planning horizon. Turnover frequencies have to be calculated for certain storage types and lower limits had to be defined. Both assumptions do only represent the real storage capacity in a limited way.

- **Process flow changes** - One possible reason for these deviations could be experiments that are being conducted. If this happens the wafer might leave its current process flow, switch to another one, and return after the experiment has finished. But still, the transports from the stopping point to the start of the experiment process flow as well as the one from the experiment and back to the original location are two wrong transports within the From-To-Matrix.
- **Reruns due to failures** - have been identified by analyzing transport chains longer than three. A transport chain is defined as the sum of single transports during one process step. Transport chains longer than three were then analyzed in detail. The main part of these chains results from failures during measuring steps. As a consequence, wafers are transported between measuring tools and buffer more than once until the measuring step could be finished on the respective tool. A helpful side effect of validating the From-To-Matrix generation is checking the real system. These transports are not wanted and worth to have a closer look at since there has to be something going wrong with them.
- **Dummy load ports** - are used when certain tools are not connected to the OHT anymore for construction reasons. To keep them running dummy load ports are used to manually deliver material to these tools. During this phase, it is important to adjust the tool delivering rules in RTD to define the right buffers since the localization of the destination tool has changed (dummy load port might be placed somewhere else than the tool it is the dummy for). For the From-To-Matrix generation, it was ok to replace the destination tool with the dummy load port. The bigger problem was the input back into the OHT since people did not always use the dummy load port but the closest manual input port which could have been a stocker, too.
- **Re-Route** - These tend to happen when a FOUP is sent to a storage location via where-next and a suitable tool becomes available during the transport. In this case, the transporting vehicle gets a new destination during execution. For transport generation, only the original start and the final destination are considered (might not influence transport generation but the vehicle routes).

For the second part, some adjustments had to be made since certain information is not available for future scenarios compared to the present. Especially, for the distribution of material to storages, the approach had to be adjusted incorporating assumptions for the turnover frequency of new storages and storage capacity needed for new tools. These were based on an algorithm for storage mapping to new tools by observing existing tool-storage assignments.

As mentioned above the layout has been checked against the forecasted transports using the FlowLogiX Monitoring and Maintenance Suite (MMS). The utilization consists of two parts: the number of loads / unloads per time unit per piece of track (link in network model) (1) as well as the number of drive-throughs (2). (1) results from the sources and sinks defined in the From-To-Matrix. (2) results from connecting the transports from source to sink using the shortest path algorithm as described earlier. The theoretical limits within the network result from a capacity model consisting of the number of loads and unloads as well as the number of drive-throughs. Suppose that the time unit is one hour which consists of 3,600 seconds, a load/unload takes 20 seconds and a drive-through only 5 seconds. Then the theoretical limit could be 180 loads/unloads or 720 drive-throughs per hour or the respective values due to the ratio of both actions. Figure 5 (routing by length) shows the 20 most utilized links and the ratio of loads/unloads and drive-throughs based on shortest paths by length. In this case, 9 sinks are over-utilized. From a possible optimization point of view, the ones with more drive-throughs are the preferable ones. Loads and unloads are defined within the From-To-Matrix. That means to reduce these values one would have to adjust the transport generation process or replace tools or buffer. Having a closer look into the buffer-selection process for sure would be the first step since this does not affect (at least not directly) tool utilization or placement. Usually, the track limitations are not part of the tool and buffer selection in the first place. But if track utilization is exceeding the maximum value, a more detailed analysis is required.



(a) Routing by length (b) Optimization to 100% capacity (c) Optimization to 80% capacity

Figure 5: Optimization of cost settings.

If over-utilized links have a significant ratio of drive-throughs it is possible to ease the situation with optimizing the cost settings. As mentioned above the iterative process tries to reduce the utilization of the most utilized links and diverts them with a minimal amount of increased transportation length. Increasing the transportation length does not necessarily mean increased transportation times since reduced traffic makes transportation faster. The results of an initial optimization run are shown in Figure 5 (b). The first optimization was carried out to lower the utilization of all over-utilized links until the 100% utilization is reached. Results show that from nine over-utilized links in the first place three could be lowered under the limit. It has to be mentioned, that five out of six remaining over-utilized links already reach the limit due to loads / unloads, and therefore routing cannot help. An important KPI to evaluate the results in the whole transportation context is the increase of average traveled length through routing optimization. This shows that the lowering of three over-utilized links comes with an average increase of 2.0% overall simulated transports. Sometimes factory managers want to have more buffer in track utilization to be more robust against system dynamics (e.g. due to batch-tool processing). In this case, it is possible to try to set the optimization goal to achieve a track utilization of 80% instead of 100%. These results show that the number of links with utilization over 80% could be lowered from 19 to 13. Again, the remaining links almost completely consist of necessary loads and unloads. Even for the 80% optimization goal, average simulated transport lengths increased by only 5.8%.

As introduced the method of from-to generation assumes a steady-state scenario for product mix. That means each process step of any given process flow is executed in a given timeframe with the same frequency (except sampling steps). To check the robustness of transport patterns the proportions of the process flows in the product mix have been scaled randomly retaining the overall load. Results have shown that the prediction of the track utilization is robust regarding changes in the product mix. Figure 6 shows the results of the original product mix and five randomized ones. The numbers of the links remain the same, the order might change due to different track utilization values.

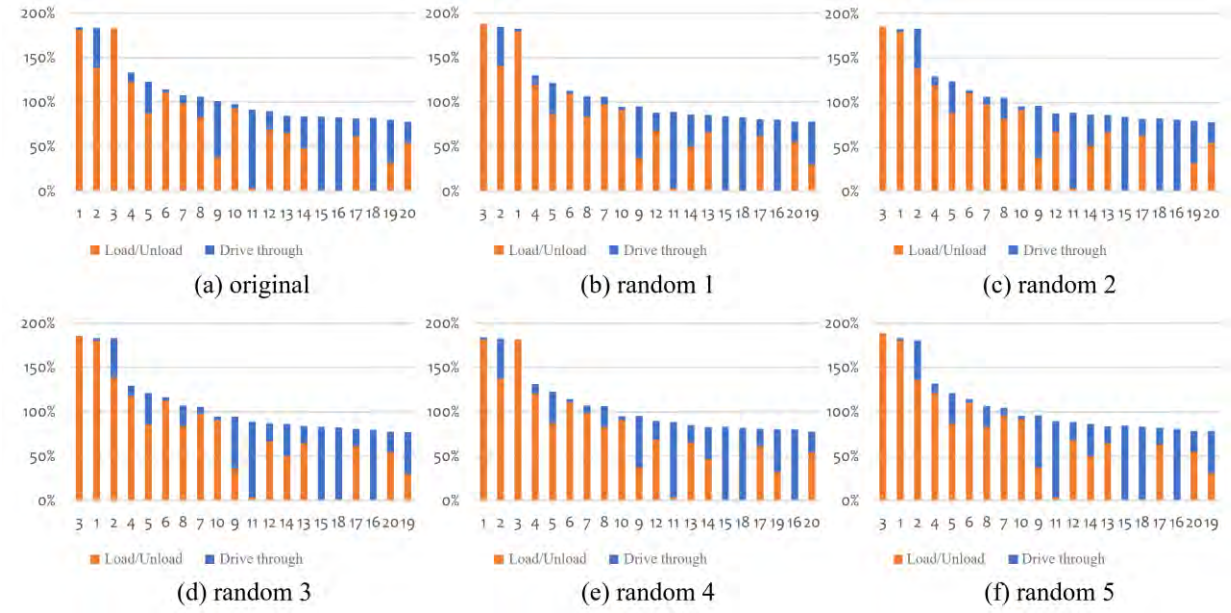


Figure 6: Link utilization for different randomized product mix scenarios.

6 CONCLUSION AND OUTLOOK

The presented approach helps to close the gap of information misfit between production and AMHS planning which has become even more important when using unified transportation systems. Compared to former approaches, the one presented here uses already existing infrastructure (RTD) and includes non-productive wafers. Therefore, any run-time information available to the dispatching team can be included in the transport generation process, too. This makes the whole process more authentic and enables a reliable prediction which parts of the AMHS might turn out to be a bottleneck for a future Fab layout and if it is going to fulfill the requirements. Additionally, it serves as a basis for discussions between production and AMHS layout planning department consequently leading to smarter decisions about where to place different (especially high throughput) tools and buffers to prevent future AMHS bottlenecks. Static Simulation is a fast way to compare different layout alternatives as well as evaluating the possible effects of different product mixes, much faster, and with less effort compared to typical dynamic Simulations. Another advantage of this approach is the deeper insight into what causes transports inside a Fab yielding possible approaches on their optimization along with an estimate of potential benefits.

There is further room for future improvements regarding the accuracy, as many simplifications are merely considered by applying a safety factor. A severe limitation is to not accurately consider empty vehicle movements. One way to include them is to use Dynamic instead of Static Simulation. Another possibility might be to add empty vehicle movements from a separate simulation approach as used in Schmaler et al. (2017) and combine these two again in the known Static Simulation.

Since a Static Simulation per se cannot consider any dynamic effects, the influence of traffic variability, e.g. caused by batch tools, has been accounted for only by an increase in the safety factor. The methodology of transport generation would need to be adjusted, by enriching the average values of the From-To-matrix with tool dependent inter-arrival-distributions. However, even Dynamic Simulation models often neglect that.

Further improvements are necessary when designing completely new Fabs without having access to reference Fabs from which experiences about test wafers and storage policies can be extracted.

References

Bartlett, K., J. Lee, A. Shabbir, G. Nemhauser, J. Sokol, and B. Na. 2014. "Congestion-aware dynamic routing in automated material handling systems". *Computers & Industrial Engineering*, 70: 176-182.

Bocaletti S., Y. Moreno, V. Latora, M. Chavez, D. U. Hwang. 2006. "Complex networks: Structure and dynamics". *Physics Reports*, 424: 175-308.

Dauzere-Péres, S., C. Yugma, A. B. Chaabane, C. M. P. Saint-Etienne, L. Rulliere and G. Lamiable. 2012. "A Study on Storage Allocation in an Automated Semiconductor Manufacturing Facility". In *Proceedings of the 12th International Material Handling Research Colloquium*, June 25th-28th, Gardanne, France, 682–692.

Gaxiola, G., D. Pabst, E. Christensen, and D. Wizelman. 2013. "Methodology to Evaluate the Impact of AMHS Design Characteristics on Operational Fab Performance". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3806-3817. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Hammel, C., T. Schmidt, and M. Schoeps. 2012. "Network optimization prior to dynamic simulation of AMHS". In *Proceedings of the 2012 Winter Simulation Conference*, editd by edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1956-1966. Piscataway, NewJersey: Institute of Electrical and Electronics Engineers, Inc. .

Hammel, C., R. Schmaler, T. Schmidt, J. Luebke, M. Schoeps, U. Horn, and M. Mosinski. 2016. "Empowering existing automated material handling systems to rising requirements" In *Proceedings of the 27th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, May 16th-19th, Saratoga Springs, USA, 87-93.

Kiba, J.-E., G. Lamiable, S. Dauzère-Pérès, and C. Yugma. 2009. "Simulation of a Full 300mm Semiconductor Manufacturing Plant with Material Handling Constraints". In *Proceedings of the 2019 Winter Simulation Conference*, edited by A. Dunkin, E. Yuecesan, M. Fu, B. Johansson, S. Jain, and J. Montoya-Torres, 1601–1609. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Lee, J. 2016. *Track Layout Accommodating Dynamic Routing in Automated Material Handling Systems*. Ph.D. thesis, Georgia Institute of Technology, GA

Lee, S., J. Lee, B. Na. 2018. „Practical Routing Algorithm Using a Congestion Monitoring System in Semiconductor Manufacturing". *Transactions on Semiconductor Manufacturing*, 31(4): 475-485.

Lin, J. T. and H. Chao-Jung. 2014. "A Simulation-Based Optimization Approach for a Semiconductor Photobay with Automated Material Handling System". *Simulation Modelling Practice and Theory*, 46: 76–100.

Murray, S., G. T. Mackulak, J. W. Fowler and T. Colvin. 2000. "A Simulation-Based Cost Modeling Methodology for Evaluation of Interbay Material Handling in a Semiconductor Wafer Fab". In *Proceedings of the 2000 Winter Simulation Conference*, edited by P. A. Fishwick, K. Kang, J. A. Joines, and R. R. Barton, 1510–1517. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Newman, M. E. J. 2003. *Routing, Flow and Capacity Design in Communication and Computer Networks.* San Francisco, CA: Morgan Kaufmann.

Rothe, J., G. Gaxiola, L. Marshall, T. Asakawa, K. Yamagata, and M. Yamamoto. 2015. "Novel Approaches to Optimizing Carrier Logistics in Semiconductor Manufacturing". *Transactions on Semiconductor Manufacturing*, 28(4): 494–501.

Schmaler, R., T. Schmidt, M. Schoeps, J. Luebke, R. Hupfer, and N. Schlaus. 2017. "Simulation Based Evaluation of Different Empty Vehicle Management Strategies with Considering Future Transport Jobs". In *Proceesings of the 2017 Winter Simulation Conference*, edited by E. H. Page, G. Wainer, V. Chan, A. D'Ambrogio, G. Zacharewicz, and N. Mustafee, 3576-3587. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Schmaler, R., C. Hammel, and C. Schubert. 2020. „AMHS Capability Assessment Based on Planned Product Mixes". In *Proceedings of the ASMC Conference*, May 4th-7th, Saratoga Springs, USA (*accepted, not published yet*)

AUTHOR BIOGRAPHIES

ROBERT SCHMALER received his Diploma in Economics and Engineering from Technische Universität Dresden in 2010. He continued to work as a Ph.D. student and a member of scientific staff at the Chair of Logistics Engineering. The work on his thesis concerning AMHS empty vehicle allocation is still in progress. Since 2018 he is with FabFlow GmbH in Dresden. His email is robert.schmaler@fabflow.de.

CHRISTIAN HAMMEL received his Diploma in Applied Mathematics (Technomathematik) in 2007 from Technische Universität Dresden. He continued to work as a PH.D. student and a member of scientific staff at the Chair of Logistics Engineering and got his Ph.D. in increasing the performance of AMHS via a network approach in 2018. In 2017 he founded FabFlow GmbH in Dresden. His email is christian.hammel@fabflow.de.

CHRISTIAN SCHUBERT graduated with a master's degree in Control Engineering from Manchester University in 2008 and received a German Diploma in 2009 from Dresden University of Technology in Mechatronics. He continued to work as a research assistant and got his Ph.D. in Multibody
Simulations in 2014. He works in Automation at Infineon Dresden. His email is Chr.Schubert@infineon.com.