Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

A STUDY ON THE IMPACT OF LOT PRIORITIES ON CYCLE TIMES IN SEMICONDUCTOR MANUFACTURING

Adrien Wartelle
Stéphane Dauzère-Pérès
Claude Yugma

Mines Saint-Étienne, Univ. Clermont Auvergne
CNRS, UMR 6158 LIMOS
880 Avenue de Mimet
13120 Gardanne, FRANCE

Quentin Christ
Renaud Roussel

STMicroelectronics
850 Rue Jean Monnet
38920 Crolles, FRANCE

ABSTRACT

This paper presents a study on the impact of lot priorities on their cycle times in a workshop within a wafer manufacturing facility using simulation. We have specifically analyzed the waiting times of lots and the associated speed up or speed down. Computational experiments were conducted using Anylogic 8 based on industrial instances from the site of Crolles of STMicroelectronics. Results indicate that a speedup of more than 300% for high-priority lots and a speed down of less than 10% are achievable when the proportion of high-priority lots remains below 10%. This study initiates a first step towards a better priority mix management, which is critical in the semiconductor manufacturing industry.

1 INTRODUCTION

In semiconductor manufacturing, silicon ingots are processed into electronic chips that can be used in modern everyday technologies. This process is arguably the most complex existing industrial process, with each wafer lot undergoing hundreds of production steps and spending several weeks or months in the manufacturing facility (wafer fab) before being finalized (May and Spanos 2006). This complexity is compounded by the necessity to produce hundreds of thousands of wafers per year for different products and clients. To meet these demands, several key performance indicators (KPIs) are used to monitor and optimize wafer fabs. The planning and scheduling of production are meant to maximize the utilization and yield of machines to make them more profitable, maximize the global throughput of the fab while minimizing its variability to ensure steady production, minimize the average and variability of the cycle times of lots, which corresponds to the time to complete all the production steps in their routes, and reduce the number of lots waiting in each work-center of the fab.

The production objectives must be considered together to address and reconcile their conflicting natures. Specifically, lots of different products and for different customers compete for the availability of machines. This competition is one of the main sources of variability, in particular in a given work-center. When a new lot arrives in a work-center or at the end of each production step performed by a machine, a decision must be made to assign a lot or a batch of lots to an available machine. This decision follows rules that are either stochastic or deterministic through a queuing discipline, such as the "First In First Out" (FIFO) rule, which selects products based on the order of their arrival in the work-center. The FIFO strategy is very common and intuitive in this case, as it is fair when lots have the same level of urgency, service time or importance. However, as lots are not identical and to take into account the differences in urgency and service times, two queuing strategies are generally used in practice and the literature:

- A production target strategy defines a set of production targets for each step of each product during certain periods. These targets are used directly for the selection and assignment of lots to machines. This strategy, called the "Production Target Dispatching Rule" (PTDR), prioritizes products that have not yet met their production targets. It assigns two temporary levels of priority class based on either the lot type or the target completion. Within a given level of priority, the FIFO rule is used to select the next lot.
- A priority class strategy assigns each lot a fixed priority class, mainly based on its cycle time target. Lots with higher priority are assigned first, and lots within the same class follow a FIFO rule. To avoid potential infinite stagnation of lots in a queue, this strategy is generally combined with others by introducing dynamic weights based on the waiting time of each lot and its initial priority class. Lots with top priority classes, often referred to as *hot lots*, are introduced by the production manager to re-balance wafer production at each work-center and stabilize the production rate of each product. *Hot lots* are also used to reduce the time to market of new products.

This study focuses on the static priority strategy, specifically on the need for production managers and decision makers to understand the global impact of the number of *hot lots* in the manufacturing system. More precisely, our focus is on examining the impact of the ratio of *hot lots* on cycle times of other "normal" lots. Although this issue was studied in the literature in the 1990s (Fronckowiak et al. 1996), it still needs to be addressed in a real-time, data-driven context with valid models that can lead to better priority assignment decisions.

The remainder of the article is structured as follows. Section 2 discusses previous works related to priority management, including general queuing theory results and results specific to semiconductor manufacturing. Section 3 describes the methodology used to conduct the computational experiments and the simulation model. Section 4 presents the results of the impact of the ratio of *hot lots* on the cycle times of both *hot lots* and "normal" lots. In Section 5, we delve into the numerical results, discuss their implications, and explore potential avenues for improvement, including refining the modeling approach. The paper concludes in the same section.

2 RELATED WORKS

This section gives an overview of the use of classical priority rules, both in Queuing Theory and in the semi-conductor industry. As illustrated later by Table 1, the introduction of a priority rule favors certain lots over others, which is crucial due to the differential in terms of required production speed. This rule requires a careful design and can serve multiple purposes:

- The improvement of the global performance, particularly in non-Markovian service systems, can be achieved by using the *Shortest Processing Time* rule. This rule reduces the overall waiting time by prioritizing clients who are quicker to serve, and thus minimizes the waiting time for other clients compared to slower ones whose waiting time would compound on all other clients, even if they arrive earlier.
- The improvement of the performance for a specific type of clients who require a given quality of service is crucial. For example, in an emergency department, patients who require resuscitation cannot afford to wait and must be treated immediately compared to all other patients.
- The improvement of fairness is the main principle addressed in the work of (Shortle et al. 2017). The authors identify several key principles:
  - *The principle of the* FIFO *(*First In First Out*) queuing discipline strategy states that an earlier arriving customer should begin service before a later arriving customer.* However, this principle does not account for differences in urgency, service type, and duration.

– *Customers with smaller service time should wait less, on average, than customers with larger service times*, which corresponds to the *Shortest Processing Time* rule. However, this principle still does not fully account for differences in urgency and service type.

This study and the following sections are focused on non preemptive fixed class queues where production steps can not be interrupted. Although not detailed here, other types of priority systems also exist such as continuous and dynamic priorities with moving weight and also preemptive queues which allow for even stricter rules where a given lot can interrupt or inhibit the production step of others lots resulting in greater speed for the priority lot concerned at the cost of lost processing work time for over lots, and thus of global productivity.

2.1 Queuing Theory

The impact of priorities on the behavior of queues and metrics for each priority class have been studied in the context of *Queuing Theory*, and in particular in the context of Markovian and half-Markovian Queuing Models (Shortle et al. 2017). These models are characterized by an arrival process that is a Poisson Process and by service times that follow an exponential distribution or (for half-Markovian models) a general distribution. This theory distinguishes between non-preemptive priority queuing systems, that do not allow the perturbation of on going services, from preemptive priority queuing systems, where a client (a lot in this paper) can take the place of another one being served if it has a greater priority.

The first observation that can be made is that the inclusion of considering priority classes only impacts the overall state probabilities, which look at every client without priority distinction, if one of the following conditions is not met:

1. No client leaves the system before being served,
2. The average service rate is the same for every class and every client,
3. The system is non-preemptive, otherwise if it is not, state probabilities do not have the same distribution and the waiting and occupancy metrics can be impacted if the interrupted work is not conserved,
4. The system is always serving a client if there is at least one, i.e. the system does not wait for clients that would be more important.

The determination of stationary state probabilities of priority models is generally difficult to obtain due to the system's complexity, which results in many rate-balance equations. Nevertheless, some results have been found such as the probabilities for priority-1 customers by (Miller 1981). Even without the full distribution of probabilities, the expectancy metrics of occupancy and waiting times can be computed using several methods including the derivation of the z-transform of the distribution, or using a direct-expected value procedure (Shortle et al. 2017).

Table 1 summarizes the main analytical expectancy formulas that have been found for Queuing Systems with non-preemptive priorities. This table uses queuing specific notations with the Kendall Notation $\mathbf{a}/\mathbf{s}/\mathbf{C}/\mathbf{K}/\mathbf{m}/\mathbf{Z}$, $\mathbf{a}$ being the probability distribution of interarrival times (M for Markovian, that is exponential, $\mathbf{G}$ for general, $\mathbf{GI}$ for general independent and $\mathbf{H_k}$ for hypo-exponential), $\mathbf{s}$ the probability distribution of service times (or processing times), $\mathbf{C}$ the number of classes, $\mathbf{K}$ the capacity of the system ($+\infty$ if no indications), $\mathbf{m}$ the size of the population if it is finite ($+\infty$ if no indications) and $\mathbf{Z}$ the queuing discipline related to the set priority rules (*FIFO* if no indications). $\mathbf{L_q}$ refers to the expected number of clients in the queue (queue length), $\mathbf{L_q^i}$ to the expected number of clients in the queue (queue length) of priority class i,$\mathbf{W_q}$ to the waiting time in the queue, $\mathbf{W_q^i}$ to the waiting time in the queue of priority class i, λ to the global arrival rate, μ to the global service rate of the system, $\rho \equiv \lambda/\mu$ to the load of the system, $\rho_\mathbf{k} \equiv \lambda_\mathbf{k}/\mu_\mathbf{k}$ to the congestion load associated to the priority class $\mathbf{k} \in \mathbf{1}..\mathbf{r}$, $\sigma_\mathbf{k} \equiv \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{k}} \rho_\mathbf{i}$ ($\sigma_\mathbf{0} \equiv \mathbf{0}$, $\sigma_\mathbf{r} \equiv \rho$) to the cumulative congestion where $\mathbf{r}$ refers to the number of classes, and $\mathbf{c}$ to the number of servers.

Table 1: Queuing expectancy metrics for non-preemptive Queuing Systems.

Model (Kendall notation)	L_q metrics	W_q metrics
M/M/1/+∞/+∞/FIFO(r=1)	$L_q = \frac{\rho^2}{1-\rho}$	$W_q = \frac{\rho}{\mu(1-\rho)}$
M/M/1/+∞/+∞/PC(r=2)	$L_q^{(1)} = \frac{\lambda_1 \rho}{\mu - \lambda_1}$	$W_q^{(1)} = \frac{\rho}{\mu - \lambda_1}$
Equal Service Rates	$L_q^{(2)} = \frac{\lambda_2 \rho}{(\mu-\lambda_1)(1-\rho)}$	$W_q^{(2)} = \frac{\rho}{(\mu-\lambda_1)(1-\rho)}$
	$L_q = \frac{\rho^2}{1-\rho}$	$W_q = \frac{\rho}{\mu(1-\rho)}$
M/H_2/1/+∞/+∞/FIFO(r=2)	$L_q^{(1)} = \frac{\lambda_1(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$	$W_q^{(1)} = \frac{(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$
Unequal Service Rates	$L_q^{(2)} = \frac{\lambda_2(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$	$W_q^{(2)} = \frac{(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$
	$L_q = \frac{\lambda(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$	$W_q = \frac{(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$
M/H_2/1/+∞/+∞/PC(r=2)	$L_q^{(1)} = \frac{\lambda_1(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho_1}$	$W_q^{(1)} = \frac{(\rho_1/\mu_1+\rho_2/\mu_2)}{1-\rho}$
Unequal Service Rates	$L_q^{(2)} = \frac{\lambda_2(\rho_1/\mu_1+\rho_2/\mu_2)}{(1-\rho_1)(1-\rho)}$	$W_q^{(2)} = \frac{(\rho_1/\mu_1+\rho_2/\mu_2)}{(1-\rho_1)(1-\rho)}$
	$L_q = L_q^{(1)} + L_q^{(2)}$	$W_q = L_q/\lambda$
M/H_k/1/+∞/+∞/PC(r=k)	$L_q^{(i)} = \frac{\lambda_i \sum_{k=1}^{r} \rho_k/\mu_k}{(1-\sigma_{i-1})(1-\sigma_i)}$	$W_q^{(i)} = \frac{\sum_{k=1}^{r} \rho_k/\mu_k}{(1-\sigma_{i-1})(1-\sigma_i)}$
Unequal Service Rates		
M/G/1/+∞/+∞/PC(r=k)	$L_q^{(i)} = \frac{\lambda_i \lambda \mathbb{E}[S^2]/2}{(1-\sigma_{i-1})(1-\sigma_i)}$	$W_q^{(i)} = \frac{\lambda \mathbb{E}[S^2]/2}{(1-\sigma_{i-1})(1-\sigma_i)}$
Unequal Service Rates		
M/M/c/+∞/+∞/PC(r=k)	$L_q^{(i)} = \frac{\lambda_i \mathbb{E}[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)}$	$W_q^{(i)} = \frac{\mathbb{E}[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)}$
Equal Service Rates		

As shown with the first two models, considering priority classes in a $M/M/1$ model does not impact the global queue length. It only introduces a factor of $(1-\rho)^{-1}$ between the mean waiting time of each class and a factor of $(\lambda_2/\lambda_1)(1-\rho)^{-1}$ between their respective queue length. As such, with a load of $\rho = 0.5$, the second priority class has to wait twice as much as the first and generates a queue length with a factor that is twice the value of the ratio between the arrival rates. Introducing unequal service rates, the expectancy metrics are similar between *PC* (Priority Classes) and *FIFO*, and there is a factor $(1-\rho_1)^{-1}$ in the priority scheme for the second priority class. With multiple classes, a factor $\frac{1-\sigma_{i-2}}{1-\sigma_i}$ appears between each class. Exact formulas for general service distribution with 1 server (M/G/1/+∞/+∞/PC(r=k)) exist with $\mathbb{E}[S^2] = \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} \mathbb{E}[S_k^2]$ (S_k is the random service time associated to class k). Exact formulas for multiple servers (M/M/c/+∞/+∞/PC(r=k)) also exist in a (fully) Markovian queue with:

$$\mathbb{E}[S_0] = \frac{(cp)^c}{c!(1-\rho)(c\mu)} \left(\sum_{n=0}^{c-1} \frac{(cp)^n}{n!} + \frac{(cp)^c}{c!(1-\rho)} \right)^{-1} \tag{1}$$

There are no exact general formulas for multi-server queues such as M/G/c/∞/∞/PC(r=k) or GI/G/c/∞/∞/PC(r=k), but there exist approximation formulas with:

$$W_q(GI/G/s) \sim (1/2)(c_a^2 + c_s^2) W_q(M/M/s) \tag{2}$$

This expression indicates that the waiting time of a GI/G/s system can be approximated with the analytical waiting time of the equivalent M/M/s system (same arrival and service rates). $\mathbf{c_a^2}$ is the squared coefficient of variation of the inter-arrival time distribution, and $\mathbf{c_s^2}$ is the squared coefficient of variation

of the service time distribution. This approximation has been extended, as a conjecture, for the metrics of each priority class with almost the same equation for $W_q^{(i)}$(GI/G/s) in (Hou and Zhao 2020):

$$W_q^{(i)}(GI/G/s) \sim (1/2)(c_a^2 + c_s^2) W_q^{(i)}(M/M/s) \tag{3}$$

2.2 Priority Mix in the Semiconductor Industry

The use of a various priorities for lots in wafer fabs is very common (Schmidt 2007). The introduction of *hot lots* allows for considerable reduction of the queuing time by skipping ahead of regular lots. In particular, *super hot lots* are sometimes used in wafer fabs to completely eliminate queuing time, either by preemptive interruptions (stopping an operation on a machine to make it available) or by preventing an operation from starting to ensure that the machine is available when a super hot lot arrive. In the case of non-fully-automated wafer fabs, these priority lots must be managed and carefully planned to ensure their delivery as soon as possible.

In addition to the potential work lost with preemptive mechanisms (Shortle et al. 2017; Schmidt 2007), the use of a priority scheme has a cycle time cost on regular lots and becomes less and less effective as the ratio of *hot lots* increases. As such, initial studies in the 1990s, from (Fronckowiak et al. 1996) with a simulation model and from (Narahari and Khan 1997) with a queuing model, have focused on measuring the impact of the ratio of *hot lots* in the priority mix. Later works have focused on determining the optimal ratios for the priority mix considering the global profit and cost of each configuration of priority class mixes (Liao et al. 2004; Kang and Lee 2007; Chang et al. 2008).

The motivation behind this study directly builds upon the existing literature. Given the criticality of the mix of lot priorities, it is critical to develop data-driven tools that can diagnose and forecast the impact of this mix on the performance of lots within each class. The present research aims to create such a tool, with the future objective of statistically validating this tool, while also facilitating the exploration and investigation of different mixes of lot priorities.

3 PROPOSED METHODS

This section presents the context of this simulation study, the models that are used and how the computational experiment were carried out.

3.1 Industrial Context

The 300mm wafer fab considered in this study is located in Crolles, France. As any other wafer fab, each lot must follow a route of operations in different work-centers of the fab. Lots must thus be assigned and scheduled on different machines.

3.2 Initial Model

In this study, the focus is on a single work-center. Data on the processing of lots in this work-center has been collected on a two-week period in 2020. This small period snapshot allows a coherent description of the fab, which changes regularly, while still including sufficient data for modelling.

The initial model implemented corresponds to a set of G/G/1 queues in a parallel network. Each queue represents a machine of the work-center, where its service time, or process time, has been modeled by fitting a positive Gamma law on the inter-departure time of products from the machine, with a filter on periods where the machine is not empty. This choice of the probability distribution has been made as it is adequate for most general service time distributions which are positive, without a long tail-behavior and can have a variability that differs greatly from its mean, contrary to an exponential distribution. As such, the model behavior can be described in three process steps:

- The lot arrival process is the first step. The arrivals have been directly taken from the data, and each lots has been randomly assigned to one of two priority classes, given a uniform distribution and a priority ratio parameter.
- The assignment process of lots to machines is the second step. A random assignment is performed using a multinomial distribution fitted on the historical assignment data of operations to machines from the subset of machines that have been able to perform this type of operations. This assignment is performed and fixed at the time of arrival of each lot in the work-center. Although it is a suboptimal assignment strategy, compared to the real behavior and also to G/G/c queuing systems where this assignment is not fixed until a lot can enter a machine, it satisfies the operation-machine associations observed and it allows an initial study of the impact of the priority mix on the waiting times of lots.
- The lot service and departure process. After waiting the availability of the assigned machine with no lots older or with a higher priority class, a lot can begin its process in the machine and departure after the process time generated, without any interruption as it is a non-preemptive system.

3.3 Design of Experiments

Based on the model described above, implemented with the Anylogic software, computational experiments were performed by varying the *hot lot* priority ratio p_{hot} from 0 to 1 (by a step of 0.01). Referring to Section 2.1 on related works, this ratio corresponds to $\frac{\lambda_1}{\lambda_1+\lambda_2}$ with $\lambda = \lambda_1 + \lambda_2$ kept fixed. The model was run 100 times for each configuration of parameters, leading to a total of 10,100 runs. For each configuration, the 17,505 arrivals of the 2-week periods are generated, and the priority of an arriving lot is randomly assigned given a binomial distribution with a probability parameter of p_{hot}. The average waiting time of the lots in the work-center is measured. The speed-up is also computed and corresponds to the ratio between the average waiting time if there was no priority lot and the actual average waiting time. The results are compared to the theoretical results of a series of M/M/1 queues using the result in Line 2, Column 2 of Table 1 (adding $+1/\mu_i$ to account for process times).

4 NUMERICAL RESULTS

4.1 Work-center Description

During the study period, 17,504 lots from 148 different products have been processed on 20 machines for 1,210 types of operations. Table 2 summarizes the load of each machine during the considered period with the mean arrival rate λ, the mean service rate μ, the congestion load $\rho = \lambda/\mu$ (showing the utilization of each machine) and the average waiting time given by the queuing model and the simulation model W_{sim}.

With almost half of the machines with $\rho > 0.8$, the set of machines of this work-center is well utilized, in particular Machine 1 which has a simulated waiting time approximately twice as large as any other machine. Although the waiting time approximation W_{MM1} differs from W_{sim}, especially for Machines 6 and 18, the trend of the waiting times is coherent and result in a global difference of 11 minutes (11%) between the global waiting time of 99 minutes from the queuing model and of 110 minutes from the simulation model. This different is small despite the fact that the mean coefficients of square variation are equal to $c_a^2 = 7.03$ and $c_s^2 = 0.66$ which, from equation (3), would suggest waiting times almost four times greater.

4.2 Numerical Results

Figures 1 and 2 show the main results of the study with the impact of a given ratio of *hot lots*. These graphs could help a production manager to decide what would be a critical upper bound ratio. For example, if the maximal speed down allowed on normal lots is set to 20%, the ratio of *hot lots* should not exceed 27%. Alternatively, if the speed up required for *hot lots* is at least 200%, then the maximal ratio should be 75%. With a mean absolute error of 0.215 (8.52%) on the speed up for *hot lots*, 0.02 (3.38%) for normal

Table 2: Machine load description.

Machine	λ (h^{-1})	μ (h^{-1})	ρ	W_{MM1} (h)	W_{sim} (h)
1	1.542	1.661	0.928	8.393	6.147 (-26.8%)
2	4.814	5.250	0.917	2.296	2.330 (+1.5%)
3	5.478	6.047	0.906	1.756	1.696 (-3.4%)
4	2.378	2.643	0.900	3.773	2.966 (-21.4%)
5	4.801	5.751	0.835	1.053	1.130 (+7.3%)
6	4.298	5.171	0.831	1.145	2.751 (+140.3%)
7	1.737	2.122	0.819	2.596	2.636 (1.5%)
8	2.071	2.557	0.810	2.056	1.841 (-10.5%)
9	1.564	1.934	0.809	2.704	3.254 (+20.3%)
10	1.737	2.230	0.779	2.031	3.328 (+63.9%)
11	4.660	6.018	0.774	0.737	0.908 (+23.2%)
12	4.221	5.478	0.771	0.796	1.077 (+35.3%)
13	1.519	2.003	0.758	2.067	2.465 (+19.3%)
14	1.673	2.210	0.757	1.863	1.628 (-12.6%)
15	1.798	2.465	0.729	1.499	1.294 (-13.7%)
16	3.689	5.379	0.686	0.592	0.680 (+14.9%)
17	1.183	1.744	0.678	1.782	1.962 (+10.1%)
18	3.404	5.783	0.589	0.420	1.015 (+141.7%)
19	1.901	3.306	0.575	0.711	1.205 (+69.5%)
20	1.635	4.347	0.376	0.369	0.380 (+3%)

lots and of 2.37 minutes (5.51%) on the waiting times for *hot lots*, 27.3 minutes (13.7%) for normal lots, the M/M/1 approximation gives a coherent view of the impact of the ratio of hots lots, especially when considering the speed up measure.

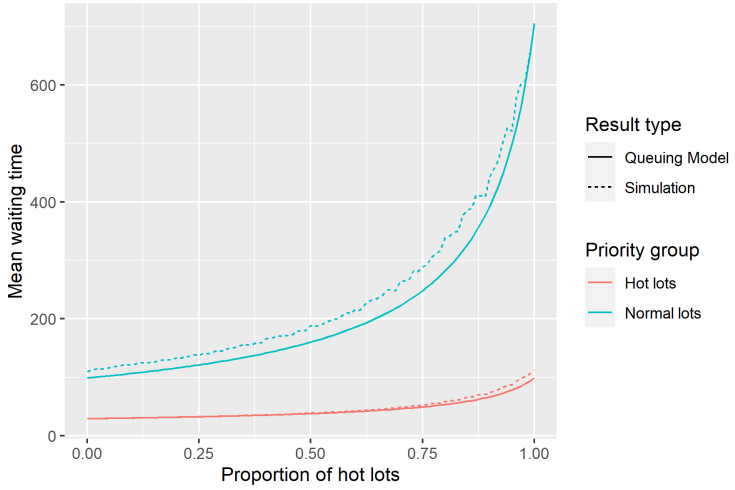

Figure 1: Average waiting time of lots depending on the priority mix.

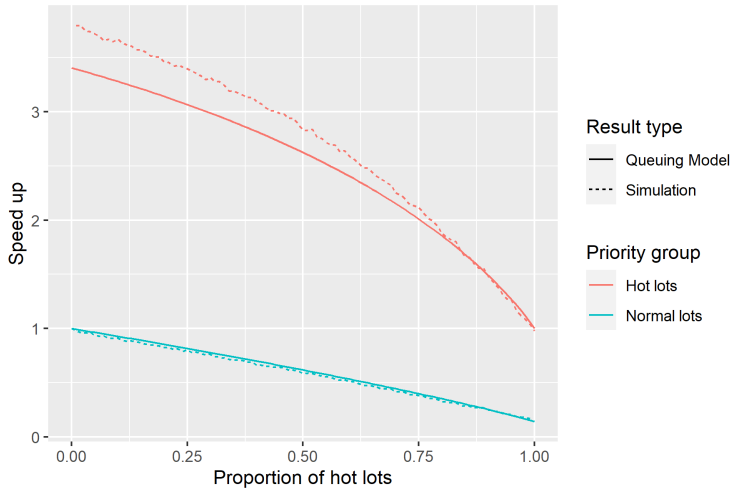

Figure 2: Speed up of lots depending on the priority mix, compared to a FIFO rule without priority classes.

5 DISCUSSION AND CONCLUSION

This study proposes an initial investigation of the impact of using a priority scheme for lots in a wafer fab work-center. With a speedup of more than 300% for *hot lots* and a slowdown of less than 10% if the ratio of *hot lots* is kept under 10% (corresponding approximately to 875 lots/week here), this strategic tool should be relevant for production managers. The impact of prioritizing some lots to speed up their cycle times should be better evaluated.

Overall, the study demonstrates that, at the scale of a work-center, it is possible to get a quick estimate of the speedup impact of a two-class priority mix using either simulation or a queuing model. Both the queuing model and the simulation model give consistent results and remain robust, contrary to what the high variability of inter-arrival would suggest. This high variability can be attributed to the regular periods when a machine is unavailable or idle due to the absence of lots that can be processed on that machine, leading to very high intervals of time that skew the variability. Still, the queuing model tends to underestimate by 13.7% the waiting time, suggesting that the high variability of inter-arrivals, when compared to the exponential case, still has an impact despite the lower service variability.

This study presents several limitations. The first is related to the extension of the results, particularly on speedup, to the real system observation. In the real system, the assignment policy promotes more efficient machine utilization and reduced queuing compared to M/M/c systems. For instance, the waiting time is 30 minutes, which is 70% less than our study's strategy. Consequently, the relative speedup of *hot lots* becomes less pronounced since the average processing time required is roughly 11 minutes.. The behaviors of the machines and lot arrivals in our study are simplified. This is due to factors such as machine maintenance leading to unavailability, setup times between two distinct consecutive operations, and other practical considerations. Furthermore, the workshop selected for our study exclusively features single-batch machines, which process lots individually. Workshops with multi-batch machines introduce added complexities and would demand additional modeling. Finally, the priority scheme used in the real system has more granularity, with six classes (Low, Standard, Medium, Hot, Super hot and ultimate). Lots with a low priority can also temporarily have a higher priority, typically if they have been waiting too long to be processed in the work-center. As such fixed priorities are not always strictly respected, notably when it can save some unnecessary setup times.

Despite its limitations, this study represents an initial stride towards enhanced priority mix management in semiconductor manufacturing. The future directions of this work can be categorized into two main streams. The first stream centers on refining the simulation model to more accurately emulate the targeted

manufacturing system, while maintaining a universal and reproducible data-driven approach. For instance, the simulation model introduced in (Anthouard et al. 2022) could be adopted, given its speed and its reliance on real-time data from a significant segment of the fab. The use of a data-based queuing discipline with the analysis of selection and assignment probabilities of lots to machine could further improve the fit with the targeted system. The second stream is the utilization of this data-driven model in the optimization of the priority mix, i.e. the assignment if priorities to lots, to improve several KPIs of a wafer fab at a work-center level and at the fab level and improve its control (see (Barhebwa-Mushamuka et al. 2023)).

REFERENCES

Anthouard, B., V. Borodin, Q. Christ, S. Dauzère-Pérès, and R. Roussel. 2022. "A Simulation-Based Approach for Operational Management of Time Constraint Tunnels in Semiconductor Manufacturing:* Topic: IE: Industrial Engineering". In *2022 33rd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 1–6. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE).

Barhebwa-Mushamuka, F., S. Dauzère-Pérès, and C. Yugma. 2023. "A Global Scheduling Approach for Cycle Time Control in Complex Manufacturing Systems". *International Journal of Production Research* 61(2):559–579.

Chang, S.-C., S.-S. Su, and K.-J. Chen. 2008. "Priority Mix Planning for Cycle Time-Differentiated Semiconductor Manufacturing Services". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 2251–2259. Piscataway, NJ: Winter Simulation Conference.

Fronckowiak, D., A. Peikert, and K. Nishinohara. 1996. "Using Discrete Event Simulation to Analyze the Impact of Job Priorities on Cycle Time in Semiconductor Manufacturing". In *IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop. Theme-Innovative Approaches to Growth in the Semiconductor Industry. ASMC 96 Proceedings*, 151–155. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE).

Hou, J., and X. Zhao. 2020. "Using a Priority Queuing Approach to Improve Emergency Department Performance". *Journal of Management Analytics* 7(1):28–43.

Kang, H.-Y., and A. H. Lee. 2007. "Priority Mix Planning for Semiconductor Fabrication by Fuzzy AHP Ranking". *Expert Systems with Applications* 32(2):560–570.

Liao, D.-Y., C.-J. Lee, and Y.-H. Lee. 2004. "A Multi-Class, Tandem Queue Model for Priority Quota Assignment in Semiconductor Manufacturing". In *Proceedings of the 5th Asia-Pacific Industrial Engineering and Management Systems Conference (APIEM2004)*, 38.1.1–38.1.14: Australian Society for Operations Research.

May, G. S., and C. J. Spanos. 2006. *Fundamentals of Semiconductor Manufacturing and Process Control*. Hoboken, NJ, USA: John Wiley & Sons.

Miller, D. R. 1981. "Computation of Steady-State Probabilities for M / M /1 Priority Queues". *Operations Research* 29(5):945–958.

Narahari, Y., and L. M. Khan. 1997. "Modeling the Effect of Hot Lots in Semiconductor Manufacturing Systems". *IEEE Transactions on Semiconductor Manufacturing* 10:185–188.

Schmidt, K. 2007. "Improving Priority Lot Cycle Times". In *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 117–121. Stresa, Italy: Institute of Electrical and Electronics Engineers.

Shortle, J. F., J. M. Thompson, D. Gross, and C. M. Harris. 2017, 1. *Fundamentals of Queueing Theory*. 5 ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.

AUTHOR BIOGRAPHIES

ADRIEN WARTELLE is a Postdoctoral Fellow at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne since January 2023 in the Manufacturing Sciences and Logistics department. He received the Ph.D degree from the University of Technology of Troyes in 2022 where he studied the modeling and simulation of congestion in healthcare systems in cooperation with the Hospital Center of Troyes. His research interests relates to the modeling and simulation of complex system in a data-driven operational research context using notably machine learning and artificial intelligence. His email address is adrien.wartelle@emse.fr.

STÉPHANE DAUZÈRE-PÉRÈS is Professor at Mines Saint-Etienne, France, and Adjunct Professor at BI Norwegian Business School, Norway. He received the Ph.D. degree from Paul Sabatier University in Toulouse, France, in 1992 and the H.D.R. from Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at M.I.T., U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. He has been Associate Professor and Professor from 1994 to 2004 at the Ecole des Mines de Nantes, France. His research interests broadly include modeling and optimization of operations at various decision levels in manufacturing and logistics, with a special emphasis on production planning and scheduling, on semiconductor manufacturing and on railway operations. He

has published more than 100 papers in international journals and contributed to more than 250 communications in national and international conferences. Stéphane Dauzère-Pérès has coordinated numerous academic and industrial research projects, including 4 European projects and more than 30 industrial (CIFRE) PhD theses, and also eight conferences. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in 2013 and the EURO award for the best theory and methodology EJOR paper in 2021. His email address is dauzere-peres@emse.fr.

CLAUDE YUGMA is Professor the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France since 2016 in Manufacturing Sciences and Logistics department. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003, and his H.D.R. from the Jean-Monnet University, Saint-Etienne, in December 2013. He was a Postdoctoral fellow at the Ecole Nationale Supérieure de Génie Industriel, Grenoble from 2003 to 2004 and from 2005 to 2006 at EMSE. He co-organized several international conferences as for example the 2013 edition of the conference Modeling and Analysis of Semiconductor Manufacturing. His research interests modeling and scheduling in semiconductor manufacturing. He has published more than 20 papers in international journals and has contributed to more than 80 communications in conferences. His email address is yugma@emse.fr.

QUENTIN CHRIST is an engineer working at STMicroelectronics. He recieved in 2020 his Ph.D degree in Industrial Engineering from the Ecole Nationale Supérieure des Mines de Saint-Etienne. His research interests include scheduling, production planning and simulation. His e-mail address is quentin.christ@st.com.

RENAUD ROUSSEL is a Scheduling and Dispatching Full Automation Expert at STMicroelectronics in Crolles (France). He has been working for more than 2 decades in the semiconductor industry in manufacturing science at the frontier between operational management, industrial engineering and data science to make the fab as efficient as possible. His email address is renaud.roussel@st.com.