applied sciences

MDPI

Article

Simulation-Based Analysis on Operational Control of Batch Processors in Wafer Fabrication

Pyung-Hoi Koo [1,*] and Rubén Ruiz [2]

[1] Department of Systems Management & Engineering, Pukyong National University, Busan 48513, Korea
[2] Instituto Tecnológico de Informática, Universitat Politècnica de València, 46021 València, Spain; rruiz@eio.upv.es
* Correspondence: phkoo@pknu.ac.kr; Tel.: +82-51-629-6485

Received: 13 August 2020; Accepted: 25 August 2020; Published: 27 August 2020

Abstract: In semiconductor wafer fabrication (wafer fab), wafers go through hundreds of process steps on a variety of processing machines for electrical circuit building operations. One of the special features in the wafer fabs is that there exist batch processors (BPs) where several wafer lots are processed at the same time as a batch. The batch processors have a significant influence on system performance because the repetitive batching and de-batching activities in a reentrant product flow system lead to non-smooth product flows with high variability. Existing research on the BP control problems has mostly focused on the local performance, such as waiting time at the BP stations. This paper attempts to examine how much BP control policies affect the system-wide behavior of the wafer fabs. A simulation model is constructed with which experiments are performed to analyze the performance of BP control rules under various production environments. Some meaningful insights on BP control decisions are identified through simulation results.

Keywords: batch processors; real-time control; dispatching; wafer fabrication; semiconductor manufacturing; system-wide performance

1. Introduction

In wafer fabrication facilities, semiconductor chips are made out of silicon wafers, thin and round slices of semiconductor material, by building electrical circuits on wafers layer by layer. Each layer requires a number of different processes, including oxidation, photolithography, etching, diffusion, deposition, ion implantation, etc. Wafers in the wafer fab move through these processes in lots generally consisting of 20–25 individual wafers. In general, a wafer lot goes through 500–700 process steps on more than one hundred machines [1]. Since semiconductor processing equipment is very expensive, it is shared by the jobs during different process steps, leading to reentrant product flow. (In this paper, the jobs refer to wafer lots.) Because of the long sequences of operations and resource sharing caused by reentrant product flow, most wafer fabs suffer from high work-in-process (WIP) inventories and long lead times (often more than one month). Since the cost of building a wafer fab is enormous, often more than ten billion dollars [2], the system capacity cannot be easily expanded. Hence, the wafer fab should be efficiently operated with a given facility capacity.

The process equipment in the wafer fabs is often classified into two types: Discrete processors (DPs) and batch processors (BPs). The DPs process wafers one at a time while the BPs process several wafer lots simultaneously as a batch. Diffusion furnace is a typical example of the batch processor where several wafer lots are placed in a reactor, which is then sealed, heated and filled with carrier gas for changes of their electrical and chemical characteristics [3]. The wafer lots arriving at the BP station are formed as a batch before being served by one of the batch processors. After a batch of wafer lots receive a process service, the batch is split into individual wafer lots before they are transferred to

a downstream station. Due to process or facility constraints, there is a limitation on the number of wafer lots in a batch that a batch processor can accommodate, which is referred to as batch capacity. Once a processing cycle begins, additional lots cannot join the batch being processed and must wait for a batch processor to become free. Due to the chemical nature of the process, it is often impossible to include the wafer lots with different recipes together in the same batch [4]. The wafer lots with the same recipe can be viewed as a job family. The job families are incompatible in that the jobs of different families cannot be processed together. In wafer fabs with reentrant characteristics, the wafers of the different processing steps need different processing recipes. Hence, even for the same wafer lots, the job family of a wafer lot depends on its current operation step. Figure 1 shows a schematic representation of a BP station with three batch processors whose batch capacity is six. In the figure, two BPs are under operation while one BP is idle because there are no job families with a full load currently waiting in the queue. (Here, a full load operation is assumed.) The processing time at the batch processors is very long compared with discrete processing times. When a BP finishes processing a batch of wafers, it releases multiple wafer lots to its downstream workstations. The bunched flow of wafer lots leads to the formation of high WIP inventories at the downstream stations, and so the manufacturing systems with BPs suffer from high flow variability. The distinct characteristics of batch processors have a significant influence on the overall wafer fab performance.

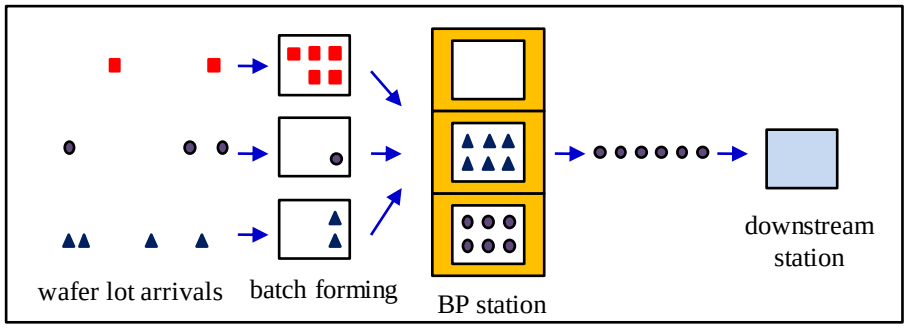


Figure 1. Schematic representation of a batch processor (BP) station with multiple batch processors.

A batch of wafers should be loaded first before they have processing service on a BP. The loading decision at a BP station is made at two instances [5]: (i) When wafer lots arrive at the BP station (a push decision); or (ii) when a batch processor becomes available (a pull decision). If a batch with a full load and a batch processor are both available, the batch is loaded and processed immediately. However, if a partial load of a batch is waiting and a batch processor is available, a loading decision should be made as to whether to load the batch immediately or to wait for more lots to arrive to form a larger batch size. Starting a batch immediately with a partial batch size undermines the BP capacity, while delayed batch loading increases the waiting time for the jobs currently waiting at the BP station, and therefore, potentially increases lead times.

Some research studies have addressed the BP control problem in semiconductor wafer fabs. Most of the BP control rules developed so far have focused on optimizing the local performance of batch processors. However, their application in multi-stage reentrant production systems may not result in as good a system-wide performance as expected [6]. For example, waiting time minimization at the BP stations may deteriorate the system-wide performance: It may lead to an unbalanced and excessive WIP level throughout the manufacturing system, resulting in increased system lead time. Hence, the decision making at batch machines should consider overall system performance, not only local performance. This paper examines the impact of BP control decisions on the system-wide performance of wafer fabs. In this paper, two system-wide performance measures are considered to examine the system performance: Throughput rate and lead times. The throughput rate may be the most important performance measure when production managers want to have a maximum throughput with given manufacturing facilities. Lead time is defined as the time needed for a wafer lot to go

all the way through a wafer fab. Lead time is widely recognized as a key performance indicator in today's lean manufacturing environments [7], whose major objective is to keep high on-time delivery with minimum WIP levels while achieving a target throughput. Little's law describes a relationship among throughput rate, lead time, and WIP inventory: $L = \lambda W$ where L is the WIP inventory, λ is the throughput rate, and W is the lead time. This law states that given a specific production rate, the lead time is proportional to the WIP level.

The objectives of this study include: (1) Identifying how much BP stations affect the behavior of the system when compared to discrete processor stations; (2) comparing the performance of the loading decisions with and without system status; and (3) examining the relationship between the local performance of the BP station and system-wide performance. A simulation model is constructed for a wafer fab, and the performance of BP operational controls is examined with simulation experiments. This paper is organized as follows: In the next section, existing BP control rules are reviewed. Section 3 describes the operational control rules for the batch processors that will be examined through simulation experiments. The simulation model and experimental results are presented in Section 4. Finally, in Section 5, conclusions are drawn, and some useful insights are discussed on BP control issues.

2. Literature Review on BP Control Rules

The BP control rules involve an event-based decision procedure for loading a batch of wafer lots on batch processors in real-time. When a batch processor completes the processing service of a batch and there exist products waiting in the queue, decisions should be made whether to start loading a batch of wafer lots right away or wait for wafer lots to arrive, and which job type to be loaded. One possible loading alternative is to load jobs immediately regardless of how many jobs exist in the waiting queue. In this case, only a single wafer lot might be loaded and processed on the batch processor. Another possible control rule is that the processing cycle is initiated only when the batch size reaches the batch capacity. In this case, the jobs which have arrived earlier should wait for upcoming jobs to form a full batch load. An in-between strategy is a threshold strategy. A commonly used threshold strategy is MBS (minimum batch size) [8]. The MBS loading strategy works as follows: Let q be the number of wafer lots waiting in the queue at a BP station, B be the predetermined MBS value (or threshold value), and C be the batch capacity. When a batch processor becomes idle, a BP control decision is initiated. If $q < B$, the BP waits until there are B jobs present at the queue—at which point, it starts serving them together. If $B \leq q \leq C$, all the q jobs are immediately loaded as a batch for having a BP process service. If $q > C$, only C wafer lots are loaded immediately for processing, and the others must wait. The determination of the optimal MBS is a main research topic in the MBS-based control strategy. A stochastic dynamic program is provided in Reference [9] to determine the optimal MBS value for a BP station. It is proved that the MBS control policy is optimal for minimizing mean waiting time when the jobs arrive with a Poisson process, and the processing times are independent and identically distributed. An MBS-based BP control problem is examined in Reference [10] for a reentrant two-stage $\delta \rightarrow \beta$ system where the first stage is a DP station with multiple discrete processors and the second is a BP station with only one batch processor. (In this paper, notations δ and β will be used for discrete and batch processes, respectively.) The discrete processors are subject to failure, which implies that the arrivals of jobs to the BP station are uncertain. Since the system is reentrant, the jobs leaving the BP station visit the DP station again. It is claimed that the control decision at the BP station affects the availability of material at the downstream station, which is especially important when the downstream DP station is a bottleneck resource. An MBS-based control algorithm is introduced in Reference [11] to determine the optimal threshold value with the objective of minimizing the number of customers in the system. It is found that at high BP traffic intensities, the system performance is relatively insensitive to the threshold value because there are always enough customers in the queue at a service completion, while at low BP traffic intensities, it is better to set the MBS value low because the time spent waiting for jobs to arrive is long. A theory-based queueing model is presented in Reference [12] to determine the threshold value for a BP station with multiple machines where multiple types of products are

processed. A closed-form formula is presented to approximate the steady-state performance measures, including lead time and WIP levels. A genetic algorithm (GA) based batching decision is proposed to find the near-optimal batch sizes for different product families.

The MBS-based strategies discussed above only consider WIPs in the queue for BP loading decisions. In modern wafer fabs with advanced shop-floor information systems, real-time manufacturing data can be collected with which smart operational controls become possible. With information on product arrivals and the state of the manufacturing system, the batch processors may be more efficiently controlled. For example, if we have information about future product arrivals at the time that a machine becomes idle, and there are wafer lots in the queue smaller than the batch capacity, we may use the information to determine whether to start the process right away or wait for future job arrivals to form a batch with more jobs. A dynamic batching heuristic, called DBH, is presented in Reference [13] for a BP station with a single BP and a single job family where information about upcoming job arrivals is considered. In DBH, the planning horizon is the BP processing time T and the number of forecasted arrivals, L, is smaller than or equal to $C - 1$ where C is the batch capacity. DBH is activated when a batch processor becomes idle, and $q \geq 1$ wafer lots are waiting in the queue. At the time epoch t_0 (current decision time), DBH first examines q against the capacity of the batch processor. If $q \geq C$, C lots are loaded and processed immediately. If $q < C$, DBH makes a batching decision to minimize the total delay time, given the forecasted job arrivals. The amount of additional delay time for products currently waiting in the queue is compared to the amount of saved delay time for the future arrivals by waiting until the arrivals occur. Figure 2 shows a waiting time change when the BP station waits for one incoming lot and load $q + 1$ lots at time epoch t_1 where t_i represents the arriving time epoch of the next ith lot after t_0. The saved wait time can be calculated as $\text{Area}_2(t_1) - \text{Area}_1(t_1)$, compared to the wait time in the immediate loading case. From Figure 2, $\text{Area}_1(t_1) = q(t_1 - t_0)$ and $\text{Area}_2(t_1) = T - (t_1 - t_0)$. The saved wait time is calculated for all the time epochs of L, and the one with the largest positive value is selected for the loading time. If all the saved wait times are negative (this means that the delayed loading leads to more waiting time), the machine starts processing jobs with the partial batch immediately.

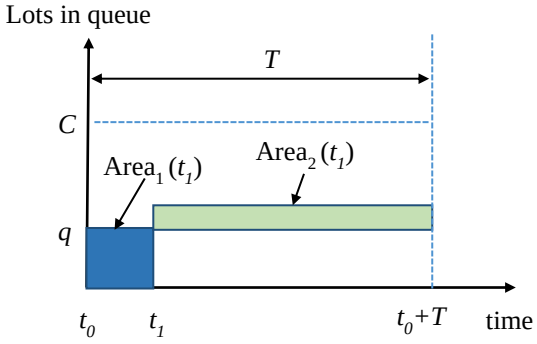


Figure 2. Changes in waiting time when the lots are loaded at time t_1. The saved wait time is $\text{Area}_2(t_1) - \text{Area}_1(t_1)$.

A BP control heuristic named the NACH (Next Arrival Control Heuristic) is presented in Reference [14] for both a single product family and multiple product families. The heuristic only considers the next job arrival and determines if it is more efficient to start the batch process immediately or at the next arrival time. When the loading is postponed until the next arrival time, the decision process is repeated at the next arrival time. It is shown that their heuristic is robust in the sense that it performs well even with prediction errors. The NACHM (NACH with multiple processors) is an extended version of NACH for a BP station with parallel batch processors [5]. Extensive review work on BP control problems with real-time job arrival data is carried out in References [15,16]. Even though BP control decisions with real-time shop-floor information mostly provide better performance than

threshold policies, most real-world wafer fabs use MBS-based policies for batching decisions [17]. This is because threshold policies are easy to implement and the future information used in look-ahead control policies is often incorrect, due to unpredictable problems, such as equipment malfunctions, product quality problems, urgent orders, etc.

Most research studies on BP control problems have considered stand-alone BP stations. A few pieces of research handle discrete processors in the downstream, upstream, and both upstream/downstream together with batch processors. A loading decision procedure is proposed in Reference [18] for a $\beta \rightarrow \delta$ network to minimize the production lead time. Experimental results show that the use of upstream and downstream information in batching decisions may lead to additional improvements at the light to moderate traffic intensity, but the improvement vanishes under high traffic conditions. A BP control heuristic for a $\delta \rightarrow \beta \rightarrow \delta$ manufacturing network is presented in Reference [19], in which the states of both upstream and downstream machines are considered. It is claimed through simulation experiments that the benefit of utilizing information about the state of an upstream discrete machine appears to be an order of magnitude larger than that of utilizing information about the state of a downstream discrete machine. An extension of the NACH policy is developed in Reference [20] for a two-stage $\beta \rightarrow \delta$ system that incorporates knowledge about future arrivals and the status of critical machines in downstream processing in the control decision process. The idea is to balance the time for the lots to spend waiting at a BP station with the time spent in the setup at a downstream DP station, thus improve the overall lead time. A rolling horizon BP control strategy is proposed in Reference [21] for a $\delta \rightarrow \beta$ network which incorporates the sequence decisions on the upstream processor through a re-sequencing approach to improve the mean lead time performance of the batch processor. Simulation experiments show that the re-sequencing approach improves the lead time performance of the $\delta \rightarrow \beta$ network as compared to the NACH and MBS-based policies, especially when the number of product types is large, and the BP traffic intensity is low or moderate.

Very few pieces of research work deal with the control problems of the batch processors from a systems perspective. The effect of multiple operational decisions, including job release, mask scheduling, and batch loading on wafer fab performance (throughput and lead time) is examined in Reference [22]. For a batch loading decision, BFQL (back and front queue leveling) heuristic is proposed in which possible starvation of the immediate downstream station is incorporated in an MBS-based loading policy. Appropriate combinations of the lot release, lot dispatching, and batching decisions are examined through simulation experiments. One possible drawback of the BFQL rule is that the control decision only considers the immediate downstream station regardless of its scheduled workload. If the scheduled workload of the downstream station is low, the system performance is not much affected by the downstream station so that the downstream status is not that important to be included in the batching decision process. The effect of dispatching rules and job release policies on system performance in wafer fabs is examined in References [23,24]. Several composite dispatching rules have been suggested, and the effect of the combination of dispatching rules and job release policies is compared through simulation experiments. However, in these studies, the effect of batch processors on system performance is not explicitly examined as only an MBS-based loading rule, or a full-batch loading rule is assumed.

To the best of our knowledge, little work has been done to use simulation analysis to examine the effect of different BP operational decisions on system-wide performance where the BP processors are explicitly considered. Our study in this paper is to examine the impact of BP control decisions on system-wide performance through simulation experiments with different BB control rules.

3. Operational Control Rules for Simulation Analysis

The control problems in wafer fab can be classified into two major problems, lot release and dispatching [24]. Lot release involves a fab-level decision that specifies when and how many wafer lots should be released to the wafer fab for the first process, while dispatching is a workstation-level decision that determines which job should be loaded next on an idle machine. There are two typical

control strategies in terms of lot release: Open-loop release and closed-loop release. The open-loop release policies release wafer lots into the fab based on static production scheduling regardless of the current system status, while the closed-loop release policies consider wafer fab status in release decision making. It is widely recognized that the performance of closed-loop release policies is better compared with that of open-loop release policies [24]. One of the simple open-loop release policies is CONTIME (CONstant TIME) where wafer lots are released into the fab at a constant time interval (for example one wafer lot is released every 2 h). With the CONTIME lot release rule, the production rate can be anticipated by the release interval, while the WIP inventories are variable depending on the operational control decisions at the workstations. A typical closed-loop release policy is CONWIP (CONstant WIP) where a constant number of WIPs are maintained in the fab [25]. In the CONWIP release rule, whenever a wafer lot leaves the wafer fab, a new wafer lot is introduced into the system. Contrary to the open-loop control systems, the throughput rates are variable depending on the operational control decisions at the workstations, while WIP levels are constantly maintained. The other widely known closed-loop release policies include the workload regulating (WR) rule [26] and the starvation avoidance (SA) rule [27] where a new lot is released to the wafer fab when the capacity load of the bottleneck station drops to a specified level. In both the WR and SA rules, the throughput is controlled by changing the critical values. Various closed-loop release rules have been proposed subsequently, which are mostly variations of the WR and SA rules [24,28].

Once the wafer lots are released into the wafer fab, they visit a number of workstations where a dispatching decision should be made as to which lot should be loaded to be processed next. Most of the existing dispatching approaches are based on priorities that are set by using product information, such as FCFS (first come first served), SRPT (shortest remaining processing time), and EDD (earliest due date), as well as system status including workload and WIP levels in downstream and/or upstream workstations. A wide variety of dispatching policies are reviewed extensively in Reference [15]. For the batch processors, the BP control decisions on batch forming and batch loading time should be made. In this paper, three control schemes, MBS-based rule, Look-upstream rule, and Look-downstream rule are examined: A control rule only with local information, a control rule with upstream information, and a control rule with downstream bottleneck station information, respectively (See Figure 3). The BP control rules are based on the existing BP control policies because our objective is not to develop new BP control rules, but investigate the impact of BP control schemes from the systems' viewpoint.

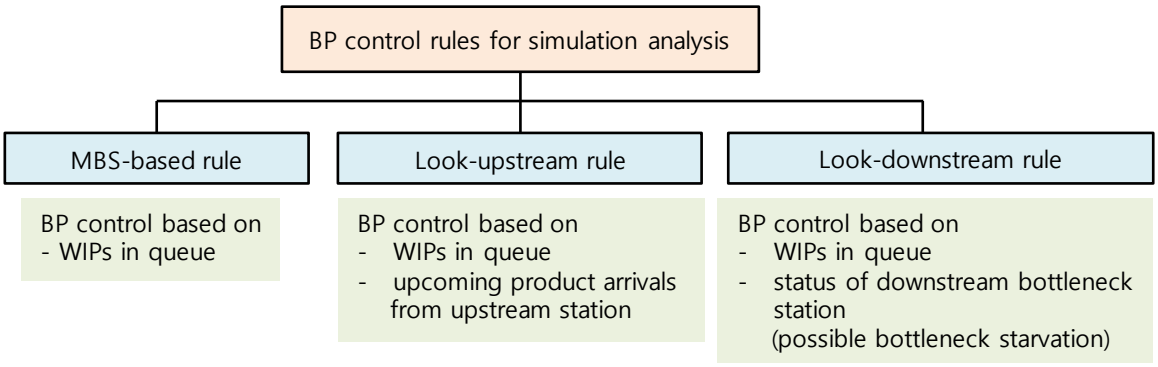


Figure 3. Three BP control rules under consideration.

(1) MBS-based rule (MBS#): Processing is initiated when the batch size of a job type is greater than or equal to a predetermined minimum batch size (MBS). The notation MBS# is used to represent the MBS rule with minimum batch size #. For example, suppose we have a batch processor whose batch capacity is six wafer lots. Then, MBS4 is an MBS rule where a batch of wafer lots is loaded when there are at least four wafer lots in the waiting queue. Two extreme cases are MBS1 and MBS6. In MBS1, whenever an idle BP finds at least one wafer lot waiting in the queue, it immediately starts loading and

processing a batch of the wafer lots. In MBS6, an idle BP starts loading only when the number of wafer lots at the queue is equal to the batch capacity (full batch loading).

(2) Look-upstream (LKUP) rule: This BP control scheme considers near-future job arrivals from the upstream stations. When a batch processor becomes idle and finds wafer lots of a product type more than the batch capacity, the wafer lots are loaded immediately. When the batch size is less than the batch capacity, the scheduler looks at the upstream stations. If a new wafer lot is expected to arrive at the BP station shortly, waiting for the new wafer lot to arrive to form a larger batch size may lead to less total waiting time at the BP station. The LKUP loading rule is similar to a look-ahead batch loading rule in References [5,13]. The arrival time of the new lot is anticipated, and the trade-off between the expected waiting time savings in the new lot and the waiting time increase in the existing wafer lots in the queue is considered. If the reduction in the waiting time is less than the increase in the waiting time, the batch is immediately loaded.

Suppose we are at the decision point, t_0, and job type i^* is selected for the next loading with batch size q. There may be jobs from the other job families waiting in the queue. When loading is delayed until the arrival time of the next job of type i^*, t_1, existing jobs waiting in the queue will experience additional delays while the new upcoming job is loaded without any waiting time. For job type i^*, the additional waiting time is $q(t_1 - t_0)$. For the other jobs in the queue, additional waiting takes place only when the jobs are loaded on the machine triggering the control decision. If there are m parallel batch processors, it is expected that the other jobs in the queue will experience delayed loading with the probability of $1/m$. Then the loading time delay (increased waiting time) can be calculated by $\frac{Q^c(t_1-t_0)}{m}$ where Q^c is the number of jobs (except job type i^*) in the BP queue. Then, the total increased waiting time, due to delayed loading is $q(t_1 - t_0) + \frac{Q^c(t_1-t_0)}{m}$. Now, we need to calculate the time saved by the delayed loading. The saved time occurs only for the incoming job. If the loading is done right away, the loading time of the next incoming job depends on the number of job types, n, and the number of batch processors. When a batch processor is available in the future, the jobs of different job types will compete with each other for the idle BP. Suppose each job type is loaded next with the same probability. Then, on average, the new job is loaded in $\frac{n+1}{2}$ loading cycles later. The interval of the machine availability (loading cycle) is $\frac{T}{m}$, where T is the BP processing time. Then, the expected waiting time is $\frac{(n+1)T}{2m}$. Now the saved wait time can be calculated by (total saved time − total increased time) as $\frac{(n+1)T}{2m} - \left[q(t_1 - t_0) + \frac{Q^c(t_1-t_0)}{m}\right]$. If the saved wait times are negative, the machine starts processing the partial batch immediately. The LKDN control rule is summarized in Figure 4.

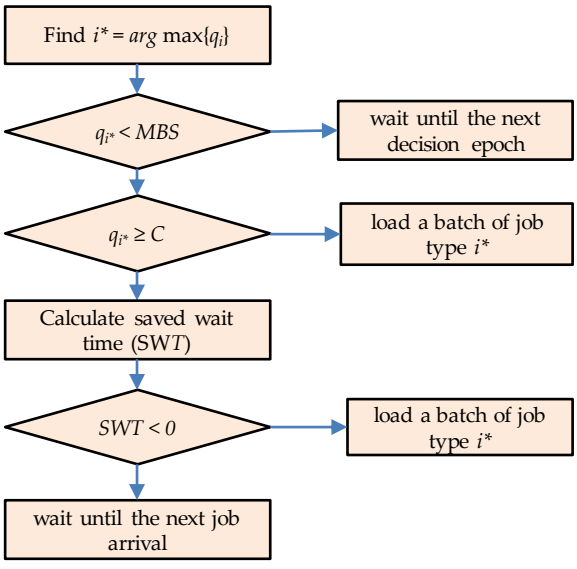


Figure 4. Decision procedure of look-upstream rule (LKUP). MBS, minimum batch size.

(3) Look-downstream (LKDN) rule: This rule considers the status of downstream machines in BP control decisions. Unlike existing control rules looking at the immediate downstream [19,22], this rule looks at a downstream bottleneck (BOT) station possibly a few steps ahead. The bottleneck station is a workstation whose facility utilization is the highest of all workstations. The detailed definition of facility utilization will be given in the subsequent section. Since the bottleneck station determines the capacity of the whole system, any idle time at the bottleneck machine undermines the system capacity. Therefore, it is important for the bottleneck machine to run without a stoppage. In this control scheme, the loading decision is made with the MBS rule in an ordinary situation. However, when the bottleneck station is expected to be idle in the near future, loading activities are initiated by forming a batch with wafer lots in the queue no matter how many wafer lots are waiting in the queue at the BP station. Here, we define the concept of virtual WIP as the work content of all jobs either in the waiting queue at the BOT station or on the way to the BOT station after visiting a BP station. Let V be the virtual WIP, p be the mean processing time per batch at the BP station and m be the number of bottleneck processors. Then, it is expected to take pV hours for the BOT station to finish all the jobs in the virtual WIP. With m BP processors in the BP station, the expected BOT starvation start time, s, is $s = pV/m$ hours when no more new jobs are additionally released from the BP station. Let t be the time required to arrive at the BOT station after being loaded on a batch processor under the condition that no waiting time is experienced on the way to the BOT station. If $s < t$, then the BOT station is expected to be idle, due to starvation. As a result, it is recommended to load the jobs immediately at the BP station to minimize the idle time of the bottleneck machines. To minimize the downtime due to starvation, some buffer WIPs are needed to absorb the flow variability. The loading decision is made based on the target BOT workload, αt, where α is the buffer factor considering system variability. An appropriate value for α (>1) is selected through preliminary simulation experiments. The LKDN control rule is summarized in Figure 5.

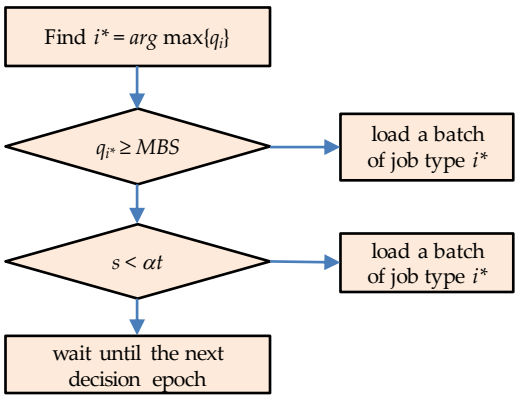


Figure 5. Decision procedure of look-downstream rule (LKDN).

4. Simulation Experiments and Results

4.1. Simulation Model Description

Simulation is a useful tool to analyze the complex and integrated system of wafer fab [29]. The simulation model in this paper is based on the wafer fab configuration from Reference [30] with a slight modification for batch processors. The wafer fab consists of 12 workstations with single or multiple processors. Even though this model is a simple version of actual wafer fabs, it contains the characteristics that a wafer fab should have, which involve reentrant product flows, batch processors, parallel machines, machine breakdown, process time variability, and different job types. The parameters describing the wafer fab are shown in Table 1, including the number of machines in each station, the number of reentrant visits, the mean process time (MPT) per wafer lot, the mean time between failure (MTBF), and the mean time to repair (MTTR). Among the 12 workstations, station 1

is a BP station while all the other stations are DP stations. The batch processors are able to process six wafer lots together as a batch. Each job follows a single product flow with 52 processing steps. The process sequence is as follows: Wafer lot release-11-12-1-9-2-3-4-5-6-7-5-1-9-2-3-4-5-10-7-5-11-12-1-9-2-3-4-5-6-7-5-1-9-2-3-4-5-10-7-5-11-12-1-9-2-3-4-5-6-7-5-8-OUT.

Table 1. Summary of The Basic Simulation Model.

Station Number	1	2	3	4	5	6	7	8	9	10	11	12
Number of Machines	2	2	3	2	3	2	2	2	2	2	3	2
Number of Reentrant Visits	5	5	5	5	10	3	5	2	5	2	3	3
MPT [1]/lot (h)	1.800	0.300	0.540	0.250	0.258	0.380	0.300	0.680	0.330	0.570	0.620	0.523
MTBF [2] (h)	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0
MTTR [3] (h)	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
Scheduled Utilization	80.0%	80.0%	95.0%	67.5%	91.0%	62.0%	80.0%	73.0%	87.5%	62.0%	67.0%	83.5%

[1] MPT, mean process time; [2] MTBF, mean time between failure; [3] MTTR, mean time to repair.

Each wafer lot visits the same station more than once (reentrant flow). The total net processing time of each wafer lot is 26.106 h on average. The processing time at each station is assumed to follow an Erlang distribution with a coefficient of variation (CV) of 0.1. The small CV value is taken because processing activities on a workstation cycle is usually computer-controlled, and as a result, typically have a low variation in the processing time. The Erlang distribution with a shape parameter k is a sum of k independent exponential variables. With a small value of CV (=1/ $\sqrt{k}$), Erlang distribution tends toward a Normal distribution by the central limit theorem, bounded on the lower side, which is a good alternative for processing times. The processing time at the BP station is independent of the number of lots in the batch and much greater than processing times at the DP stations. Each machine is individually subject to random downtime as wafer fab is subject to many sources of variability. The time between failure (uptime) and the time to repair (downtime) is randomly generated from an exponential distribution with given mean values.

In most wafer fabs, the photolithography, station 3 in our case, is a bottleneck station with the highest facility utilization which determines the system capacity. In this study, facility utilization (u) is defined as:

$$utilization(u) = \frac{MTTR}{MTBF+MTTR} + \frac{(production\ rate)(number\ of\ operation\ steps\ visited)(MPT)}{(number\ of\ processors)(batch\ capacity)} \quad (1)$$

For example, let us take station 1 with two batch processors in Table 1. Suppose 24 wafer lots are produced daily, i.e., one wafer lot/hour. Every wafer lot visits this station five times before leaving the wafer fab, and the processing time for each visit is 1.8 h. Each processor in station 1 is subject to downtime of 2 h every 38 h on average. Then, the utilization of station 1 is calculated by $\frac{2}{38+2} + \frac{1\times5\times1.8}{2\times6} = 0.05 + 0.75 = 0.8$, i.e., 80.0%.

The simulation model is written using the ARENA modeling tool, version 14.50.00, while specialized requirements are handled by using Visual Basic for Applications (VBA) subroutines. Simulation runs are done on a laptop computer with an Intel Core i7-4500U Processor @ 1.8 GHz. Experimental runs have been replicated ten times with ten different random seeds for each manufacturing scenario to reduce the effect of randomness on the performance. Each simulation run was made for the system under a one-year 24-h-a-day operation. To obtain meaningful steady-state results from the simulation runs, statistics on the initial transient period (the first three months) of each run were excluded from the analysis. After the warm-up period, statistical data is collected through simulation runs for nine months (270 days, 6480 h).

Since this paper focuses on the batch processors, commonly used control rules are applied for lot release decisions and dispatching decisions for DP stations. For lot release rules, two input control rules, an open-loop CONTIME, and a closed-loop CONWIP are applied. The CONTIME rule is selected because the production rate is anticipated with certainty, while CONWIP is selected because it is an easy-to-use closed-loop control rule widely used in industries. The inter-arrival time of the CONTIME rule is one hour, leading to a 95% utilization of the bottleneck station as in Reference [31]. (See Table 1 for the utilization for every workstation.) The number of lots maintained in the fab for the closed-loop CONWIP rule is 70, which is chosen through preliminary experiments so that the average throughput rate is roughly the same as the average throughput rate of the CONTIME lot release cases. For dispatching in the discrete processing stations, the FCFS rule is used because it is easy to apply and widely used in industries.

4.2. Results from Simulation Experiments

Two performance measures are considered to examine the system performance: Throughput rate and lead times. For the CONTIME input rule, since the production rate is determined by the job inter-arrival time, the system lead time is used as a performance measure. On the other hand, for the CONWIP input rule, since the WIPs in the system are maintained at a constant value, the production rate is used as a performance measure.

4.2.1. Effect of BP Utilization on System Performance

Simulation experiments have been performed under the CONTIME lot release condition to see the effect of batch processors on lead time. The utilization for each station remains the same as seen in Table 1, except for the BP station that has three different utilizations. To have different utilizations, the MPT for station 1 is adjusted by using Equation (1). For example, to have a utilization of 70% at station 1, MPT $= \left(0.70 - \frac{2}{38+2}\right)\left(\frac{2\times6}{1\times5}\right) = 1.56$ h. The reason why we change the MPT instead of changing the arrival rate to have different BP utilization levels is that we want to have the utilization of the other stations remain the same. Figure 6 shows the waiting time at each station. It is seen that station 1, the BP station, has the longest waiting time even with a low BP utilization, such as 60%. The waiting time at the BP station is even higher than the waiting time at the bottleneck station, station 3 with 95% utilization. The result indicates that to reduce production lead time, operation managers should pay close attention to the control of batch processors.

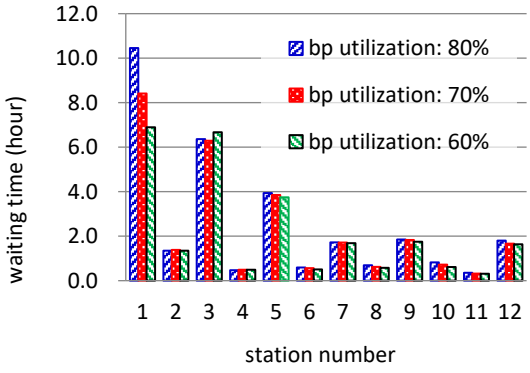


Figure 6. Waiting time at each station with different BP utilization levels.

To examine the impact of the BP utilization and DP utilization on system performance, we have performed experiments for the systems with varying utilization levels for station 1 (BP) and station 7 (DP). Here, to see the throughput change over different utilization levels, the closed-loop lot release rule, CONWIP, is used. The experimental results are given in Figure 7. When we examine the performance over varying BP utilization levels, the throughput decreases as the BP utilization increases. The decrease rate is larger in higher BP utilization levels than in lower BP utilization levels. When BP

utilization is 80%, the system produces 6492 wafer lots, which is 94.9% of the capacity. (Note that the system capacity for nine months is 6840 wafer lots.) The lead time also increases slowly with the lower BP utilization, but increases rapidly with the higher BP utilization. When we compare the impact of BDs and DPs, we can find that the change of the throughput and lead time under different utilization levels are smaller in the DP case than in the BP case. The BP station is more sensitive than the DP station in terms of the effect of the utilization levels on the system performance. For example, the system with 60% BP utilization and 80% DP utilization produces 6656 wafer lots, while the system with 80% BP utilization and 60% DP utilization produces 6537 wafer lots (−119 lots). The results indicate that the batch processor should have more planned excessive capacity than the discrete processors. This is quite an important result for practitioners.

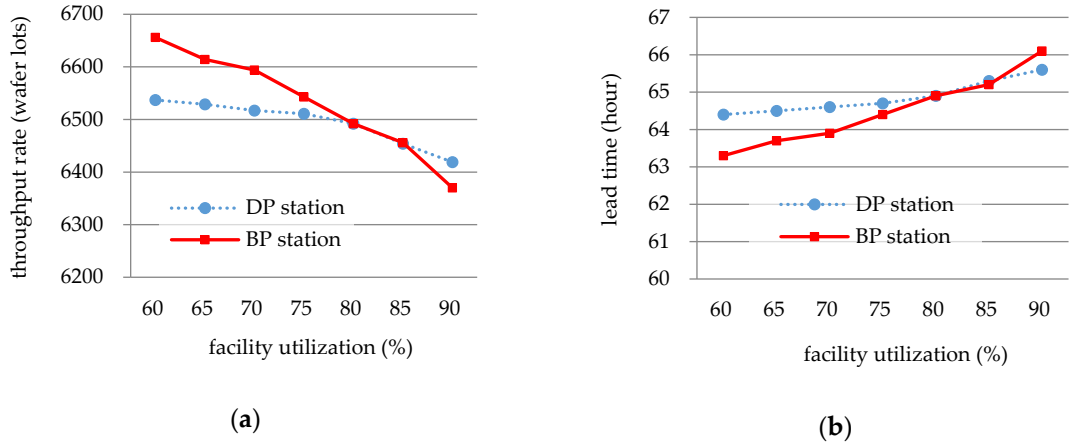
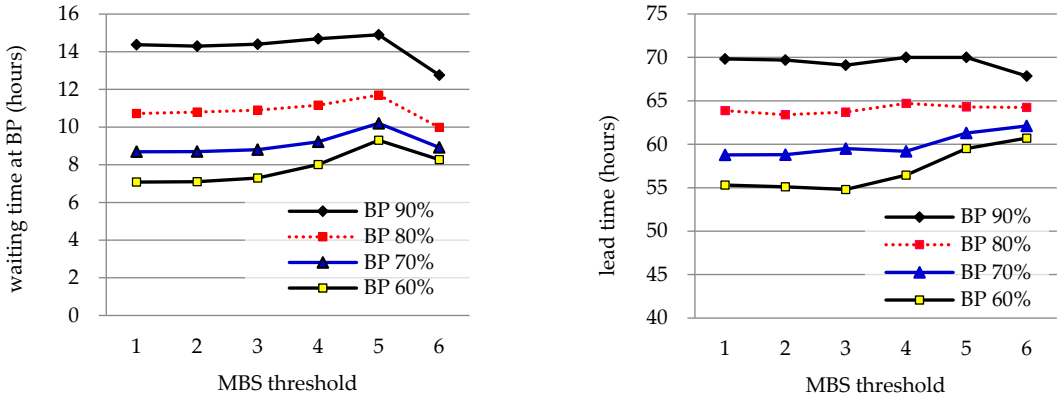


Figure 7. Performance compared over different BP and DP utilization levels in terms of (**a**) throughput rate per nine months and (**b**) system lead time.

4.2.2. Effect of the MBS Size on System Performance

In MBS-based control schemes, the determination of the MBS threshold value is a primary decision. It is known from previous research studies that the performance of the MBS threshold values depends on the BP utilization levels. To examine the combined impact of MBS threshold values and BP utilization levels, simulation experiments are carried out with six different MBS threshold levels with four different BP utilizations. Here, to see the changes in waiting time and lead time over different BP utilization levels, the open-loop lot release rule, CONTIME, is used. The experimental results are shown in Figure 8. It is seen that overall, lower MBS threshold values provide less system lead time, as well as less waiting time at batch processors. This is especially true regarding lower BP utilization levels, such as 60% or 70%. However, when the BP utilization is high, such as 80% or 90%, the MBS values have little effect on the performance. This is because the batch size is large, with a high BP utilization level regardless of the MBS threshold values.



(**a**) (**b**)

Figure 8. Performance of different MBS levels batch sizes in the MBS control in terms of (**a**) waiting time at BP station and (**b**) system lead time.

One interesting finding is that as MBS levels increase from one to five, BP waiting time tends to increase slowly. However, with an MBS threshold of six, BP waiting time drops sharply compared to an MBS threshold of five. This is arguably due to the reentrant characteristics. Note that in the system under consideration, the batch capacity of a BP is six. Hence, in MBS6, only six wafer lots are formed as a batch to have a process service and are released at the same time to a downstream workstation. They visit several discrete workstations before arriving at the next batch processor where six lots are again required to form a batch. Their arrival at the next BP station should be within a relatively small period, so less waiting time is needed to form a batch at the BP station. When the BP utilization is high, the BP waiting time drops more sharply. In the case of a BP utilization of 80% and 90%, MBS6 provides the lowest waiting time at the BP station. However, from Figure 8b, it is seen that the sharp decrease in BP waiting time under MBS6 does not lead to less system lead time. The lead time is increased with MBS6 compared to MBS5 except in the case of high BP utilization. The results indicate that less BP waiting time at the local station does not always lead to less system lead time. Industries often utilize a full load control policy in the BP station. This may be a good idea when the BP station is under a heavy workload. However, in most cases, the BP station is not critically constrained in its capacity. In this case, a lower MBS threshold is recommended for better system performance.

4.2.3. Performance of Look Upstream Control Rule (LKUP)

This rule attempts to reduce the waiting time at the batch processing station by considering jobs incoming shortly. The loading with a partial batch is delayed until a new job arrives shortly if this delay is expected to lead to less waiting time. Table 2 compares the waiting time at the BP station when MBS1 and LKUP rule are utilized for BP loading, under the CONTIME job release environment. It is seen that the LKUP rule provides consistently less waiting time at the BP station than the base case regardless of BP utilization levels.

Table 2. Comparison of the Performance of LKUP (Look-Upstream) and MBS1 (Minimum Batch Size 1).

BP Utilization	Waiting Time at BP Station (h)			System Lead Time (h)		
	MBS1	LKUP	% Difference	MBS1	LKUP	% Difference
60%	7.08	6.79	−4.1%	55.30	55.61	0.6%
65%	7.82	7.60	−2.8%	57.17	57.08	−0.2%
70%	8.69	8.40	−3.3%	58.78	58.69	−0.2%
75%	9.61	9.24	−3.9%	61.71	60.23	−2.4%
80%	10.72	10.39	−3.1%	63.87	62.58	−2.0%
85%	12.30	11.74	−4.6%	65.89	64.78	−1.7%
90%	14.37	14.16	−1.5%	69.82	68.94	−1.3%

Production operation managers are more likely interested in such global performance measures as system lead time and throughput than the local performance measures, such as waiting time at the BP station. From Table 2, it is found that the LKUP performs better than MBS1 when the BP utilization is high. However, MBS1 and LKUP rules provide almost the same lead time with lower BP utilization around 60% or 70% (Note that in these BP utilization ranges, The LKUP performs better than MBS1.) The simulation results indicate that the locally good control strategy may not lead to good system performance. This insight into the control scheme may be especially true when the product flow has a reentrant characteristic. To examine the effect of reentrant flow on the performance, we have carried out some simulation experiments with scenarios with short serial product flow. To have a serial flow system, the original product flow requiring 52 operations is divided into five independent jobs where each job visits the bottleneck machine and BP station only once, respectively. Table 3 compares the BP waiting time and lead time for the serial and reentrant product flows. Here, the lead time for the serial product flow is multiplied by five to make a fair comparison. It is seen that the local waiting time is reduced for both reentrant and serial product flow environments with the LKUP control rule. However, the system-level lead time remains almost unchanged in the reentrant flow case, whereas the lead time is reduced in the serial product flow case.

Table 3. Performance of LKUP (Look-Upstream) for the Serial and Reentrant Product Flows.

Flow Type	Waiting Time at BP Station (h)			System Lead Time (h)		
	MBS1	LKUP	% Difference	MBS1	LKUP	% Difference
serial flow	10.4	10.2	−2.4%	60.3	59.6	−1.2%
reentrant flow	8.7	8.4	−3.3%	58.8	58.7	−0.1%

4.2.4. Performance of Look Downstream Control Rule (LKDN)

The LKDN rule loads the wafer lots even with a smaller batch size than MBS level when the downstream bottleneck workload is less than the predefined target workload, 35 wafers in this case. The target workload is obtained from preliminary experiments. For the benchmark scenario, the MBS6 rule is used. Figure 9a,b show the throughput rate and lead time under the LKDN control rule and the MBS6, under the CONWIP job release environment. It is seen that the LKDN rule gives better performance than the MBS6 with more throughput rate and less lead time, especially with lower BP utilization levels. Since most wafer fabs keep the BP utilization less than 80%, the LKDN rule may be a good alternative to the MBS6. The high performance of LKDN can be realized by utilizing the bottleneck station (BOT) with less starvation downtime. Figure 9c compares the performance of LKDN and MBS6 in terms of the BOT utilization over different BP utilizations. The figure shows that the LKDN provides higher BOT capacity utilization than the MBS6, especially when the BP utilization is

not too high. It should be noted that an increase in BOT capacity utilization directly leads to in greater throughput, which means a great deal in the capital-intensive wafer fab.

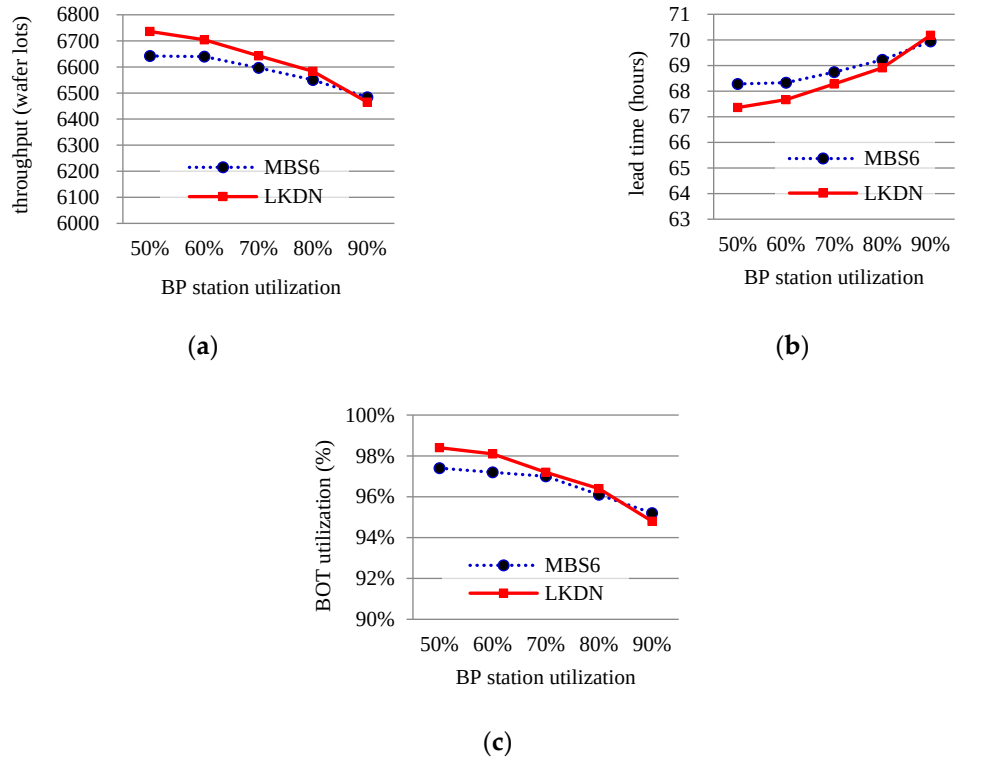


Figure 9. Performance of LKDN (Look-Downstream) and MBS6 (Minimum Batch Size 6) over different BP utilization levels in terms of (**a**) throughput, (**b**) production lead time, and (**c**) BOT (bottleneck) station utilization.

5. Conclusions

This paper attempts to provide managerial insights about the operational control policies at batch processing stations in wafer fabs. Batch processors have distinct characteristics different from discrete processors: The long batch processing time and non-smooth product flow caused by repetitive batching and splitting increase the flow variability, resulting in long lead times. Most previous studies on batch processors focus on a stand-alone BP station. Not much study has been done for batch processor control from a systems perspective. We have constructed a simulation model for a wafer fab and performed experiments with a variety of operational environments. From simulation studies, we have collected some interesting findings as follows:

(1) Batch processors are more sensitive than discrete processors in terms of capacity changes. A small change in capacity in the batch processors has more effect on the system performance than discrete processors. As a result, batch processors should have more excess capacity than discrete processors.

(2) When the BP utilization is not very high, lower MBS threshold values perform better. In the case of very high BP utilization levels, the selection of the MBS level has little effect on the performance of the system. This claim is parallel to the same results studied in References [11,32] in which only a stand-alone BP station is considered.

(3) In the wafer fab with the reentrant product flow, the batch size of a full load may decrease the waiting time at the BP stations. However, it is seen that the decrease in the BP waiting time does not guarantee a good system-level performance. Our experiments show that the full load batching policy provides longer lead times than the MBS policy with lower threshold values,

especially in the cases of moderate and lower BP utilization. The result is contradictory to the common belief about batch size determination in industries where BP operation with a full load is widely used. It is believed that the result is one of the major contributions of this paper.

(4) When a look-upstream control policy is used, it leads to the lower waiting time at the BP station. However, simulation experiments show that the look-upstream control policy does not lead to less system lead time. When the manufacturing system is characterized as a reentrant flow, the control schemes considering future job arrivals may not provide better system performance as expected. When the benefit of the look-upstream strategy is minimal from the systems perspective, it may be a good idea to apply simple rules like MBS without considering upcoming job arrivals.

(5) The look-downstream control policy performs better for the system, especially with low or moderate BP utilization levels than the full load batching policy. The result is contradictory to the argument from Reference [19] insisting the control decision considering the state of downstream DP station has little effect on the system performance. The contradictory result is arguably due to the manufacturing environment with which the downstream DP station is a highly utilized bottleneck with 95% utilization in our system, while the utilization of downstream DP station is not that high, i.e., 80% in Reference [19]. When the utilization of the BP station is high, LKDN and MBS provide almost the same performance.

The study in this paper may be improved on different levels. Firstly, a variety of testbeds may be modeled to examine the impact of batch processors with more scenarios. The limitation of our study in this paper is that we only consider a specific wafer fab model, and hence, the experimental analysis is done for a limited manufacturing case. Research work is invited to perform simulation experiments with a variety of wafer fab models, especially within the real-life wafer fab settings to examine the system behavior affected by batch processors. Secondly, some manufacturing settings not covered in this paper, such as sequence-dependent BP setup times, waiting time constraints, BP stations with different batch capacity, multiple product types with different product flows, and orders with due dates, are interesting issues to be addressed. When due dates are involved for each job (order), performance measures, such as tardiness and lateness may be more important than the lead time, the performance measure in this paper. In this case, new BP control rules need to be devised to improve upon the presented rules. Finally, it would be interesting to examine how lot release, DP dispatching (for bottleneck and non-bottleneck stations) and BP control policies interact with each other and how the optimal combination of the control policies is selected under different manufacturing settings.

Author Contributions: Conceptualization, P.-H.K.; methodology, P.-H.K. and R.R.; software, P.-H.K.; validation, P.-H.K. and R.R.; writing—original draft preparation, P.-H.K.; writing—review and editing, R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Pukyong National University Research Abroad Fund (C-D-2016-0843).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, L.C.; Chu, P.-C.; Lin, S.-Y. Impact of capacity fluctuation on throughput performance for semiconductor wafer fabrication. *Robot. Comput. Integr. Manuf.* **2019**, *55*, 208–216. [CrossRef]
2. Ham, M. Integer programming-based real-time dispatching (i-RTD) heuristic for wet-etch station at wafer fabrication. *Int. J. Prod. Res.* **2012**, *50*, 2809–2822. [CrossRef]
3. Mathirajan, M.; Sivakumar, A. A literature review, classification and simple meta-analysis on scheduling of batch processors in semiconductor. *Int. J. Adv. Manuf. Technol.* **2006**, *26*, 990–1001. [CrossRef]
4. Koo, P.; Mansoer, P. A look-ahead control strategy at parallel batch processors with multiple product types. *ICIC Express Lett. Part B Appl.* **2015**, *6*, 3197–3203.
5. Fowler, J.W.; Hogg, G.L.; Phillips, D.T. Control of multiproduct bulk service diffusion/oxidation processes. Part 2: Multiple servers. *IIE Trans.* **2000**, *32*, 167–176. [CrossRef]

6. van der Zee, D. Adaptive scheduling of batch servers in flow shops. *Int. J. Prod. Res.* **2002**, *40*, 2811–2833. [CrossRef]
7. Wang, J.; Zheng, P.; Zhang, J. Big data analytics for cycle time related feature selection in the semiconductor wafer fabrication system. *Comput. Ind. Eng.* **2020**, *143*, 106362. [CrossRef]
8. Neuts, M.F. A general class of bulk queues with Poisson input. *Ann. Math. Stat.* **1967**, *38*, 759–770. [CrossRef]
9. Deb, R.K.; Serfozo, R.F. Optimal control of batch service queues. *Adv. Appl. Probab.* **1973**, *5*, 340–361. [CrossRef]
10. Gurnani, H.; Anupindi, R.; Akella, R. Control of batch processing systems in semiconductor wafer fabrication facilities. *IEEE Trans. Semicond. Manuf.* **1992**, *5*, 319–328. [CrossRef]
11. Avramidis, A.N.; Healy, K.J.; Uzsoy, R. Control of a batch-processing machine: A computational approach. *Int. J. Prod. Res.* **1998**, *36*, 3167–3181. [CrossRef]
12. Fowler, J.W.; Phojanamongkolkij, N.; Cochran, J.K.; Montgomery, D.C. Optimal batching in a wafer fabrication facility using a multiproduct G/G/c model with batch processing. *Int. J. Prod. Res.* **2002**, *40*, 275–292. [CrossRef]
13. Glassey, C.; Weng, W. Dynamic batching heuristic for simultaneous processing. *IEEE Trans. Semicond. Manuf.* **1991**, *4*, 77–82. [CrossRef]
14. Fowler, J.W.; Phillips, D.T.; Hogg, G.L. Real-time control of multiproduct bulk-service semiconductor manufacturing processes. *IEEE Trans. Semicond. Manuf.* **1992**, *5*, 158–163. [CrossRef]
15. Sarin, S.C.; Varadarajan, A.; Wang, L. A survey of dispatching rules for operational control in wafer fabrication. *Prod. Plan. Control* **2011**, *22*, 4–24. [CrossRef]
16. Koo, P.; Moon, D. A review on control strategies of batch processing machines in semiconductor manufacturing. *IFAC Pap. Ser. Title Manuf. Model. Manag. Control* **2013**, *7*, 1690–1695. [CrossRef]
17. Leachman, R.C.; Kang, J.; Lin, V. SLIM: Short cycle time and low inventory in manufacturing at Samsung Electronics. *Interfaces* **2002**, *32*, 61–77. [CrossRef]
18. Robinson, J.K.; Fowler, J.W.; Bard, J.F. The use of upstream and downstream information in scheduling semiconductor batch operations. *Int. J. Prod. Res.* **1995**, *33*, 1849–1869. [CrossRef]
19. Neale, J.; Duenyas, I. Control of manufacturing networks which contain a batch processing machine. *IIE Trans.* **2000**, *32*, 1027–1041. [CrossRef]
20. Solomon, L.; Fowler, J.W.; Pfund, M.; Jensen, P.H. The inclusion of future arrivals and downstream setups into wafer fabrication batch processing decisions. *J. Electron. Manuf.* **2002**, *11*, 149–159. [CrossRef]
21. Cerekci, A.; Banerjee, A. Effect of upstream re-sequencing in controlling cycle time performance of batch processors. *Comput. Ind. Eng.* **2015**, *88*, 206–216. [CrossRef]
22. Kim, Y.; Lee, D.H.; Kim, J.U. A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities. *J. Manuf. Syst.* **1998**, *17*, 107–117.
23. Bahaji, N.; Kuhl, M.E. A simulation study of new multi-objective composite dispatching rules, CONWIP, and push lot release in semiconductor fabrication. *Int. J. Prod. Res.* **2008**, *46*, 3801–3824. [CrossRef]
24. Li, Y.; Jiang, Z.; Jia, W. An integrated release and dispatch policy for semiconductor wafer fabrication. *Int. J. Prod. Res.* **2014**, *52*, 2275–2292. [CrossRef]
25. Spearman, M.L.; Woodruff, D.L.; Hopp, W.J. CONWIP: A pull alternative to Kanban. *Int. J. Prod. Res.* **1990**, *28*, 879–894. [CrossRef]
26. Wein, L.W. Scheduling semiconductor wafer fabrication. *IEEE Trans. Semicond. Manuf.* **1998**, *1*, 115–130. [CrossRef]
27. Glassey, C.R.; Resende, M.G.C. Closed-loop job release control for VLSI circuit manufacturing. *IEEE Trans. Semicond. Manuf.* **1998**, *1*, 36–46. [CrossRef]
28. Qi, C.; Sivakumar, A.I.; Gershwin, S.B. An effective new job release control methodology. *Int. J. Prod. Res.* **2009**, *47*, 703–731. [CrossRef]
29. Fowler, J.W.; Monch, L. Discreet-event simulation for semiconductor wafer fabrication facilities: A tutorial. *Int. J. Ind. Eng. Theoryappl. Pr.* **2015**, *22*, 661–682.
30. El-Khouly, I.; El-Kilany, K.S.; Young, P. A simulation study of lot flow control in wafer fabrication facilities. In Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management, Istanbul, Turkey, 3–6 July 2012; pp. 2240–2249.

31. Kim, Y.; Kim, G.; Choi, B.; Kim, H. Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. *IEEE Trans. Robot. Autom.* **2001**, *17*, 589–598.
32. Akcali, E.; Uzsoy, R.; Hiscock, D.G.; Moser, A.L.; Teyner, T.J. Alternative loading and dispatching policies for furnace operations in semiconductor manufacturing: A comparison by simulation. In Proceedings of the 2000 Winter Simulation Conference, Orlando, FL, USA, 10–13 December 2000; pp. 1428–1435.