 The definitive version of this paper has been published in Winter Simulation Conference '99, 1999.

CONLOAD – A NEW LOT RELEASE RULE FOR SEMICONDUCTOR WAFER FABS

Oliver Rose

Institute of Computer Science
University of Würzburg
97074 Würzburg, GERMANY

ABSTRACT

In this paper, we present CONLOAD (CONstant LOAD), a new lot release rule for wafer fabs. It was developed to overcome some performance problems of traditional lot release rules like CONWIP or Workload Regulation during product mix changes. We show that CONLOAD outperforms CONWIP and Workload Regulation with respect to keeping the bottleneck utilization at a desired level and to provide a smooth evolution of the WIP.

1 INTRODUCTION

Product mix changes are an issue for all kinds of semiconductor fabrication facilities. They occur quite often in ASIC fabs because there is only a rather limited amount of wafers required for each product. They take place in memory or processor fabs due to modifications of chips of a certain technology or due to changes in production technology. Product mixes are not changed instantaneously. For instance, if one product is replaced by a new one, it takes weeks until all lots of the old product leave the fab even though only lots of the new product are released to fab. Because the new lot may have a different number of layers, a different number of machines to manufacture a single layer, or a different processing time at the bottleneck workcenter, the change in product mix will affect the fab performance in terms of WIP (work in progress) and cycle times. Only very little is known about the fab behavior during the transient phase induced by product changes. There could be temporal overload leading to large inventories and cycle time, or there could be temporal drops in load leading to capacity losses.

To run the fab smoothly during product mix changes, lot release rules can be applied. *Pull* rules, like CONWIP (Hopp and Spearman 1996) or Workload Regulation (Lawton, Drake, Henderson, Wein, Whitney, and Zuanich 1990; Wein 1988), may be used to draw fresh lots from an inventory buffer based on the current fab status, e.g., in terms of WIP or bottleneck utilization. By means of these rules, it is attempted to avoid overload and to smooth the stream of lots flowing into the fab. In contrast, *push* rules release lots to the fab without taking into account the current status of the fab. During the course of our study, it turned out that both CONWIP and Workload Regulation were not capable to avoid overload because of their lack in tracking the current load situation of the fab accurately enough.

We therefore developed a new lot release rule which we termed CONLOAD (CONstant LOAD). In contrast to the other two rules, CONLOAD takes into consideration how much load is added to a single machine or a group of machines by a particular lot to decide on releasing this lot into the fab or not.

For the comparison of the performance of the three rules CONWIP, Workload Regulation, and CONLOAD, we have to take performance measures like WIP or bottleneck utilization over time. To carry out simulations with real fab models consisting of several hundred machines would take a considerable amount of simulation time and the generalization of the results would be questionable because the results might vary for different fab models. Therefore, we use a simple fab consisting of a detailed model of the bottleneck workcenter and delay units representing the remaining machines of the fab. This model already proved to be useful in analyzing the behavior of wafer fabs in (Rose 1998) and (Rose 1999).

The paper is organized as follows. The considered lot release rules are presented in Section 2. In Section 3 we outline the simple fab model and simulation details. The comparison of the performance of the lot release rules for different scenarios is provided in Section 4.

2 LOT RELEASE RULES

A number of studies show that wafer fabs under pull regimes or closed loop control outperform those with traditional push or uniform release rules with respect to a number of performance measures like WIP or average and variance of cycle times. The main reason for this positive effect of pull rules is the dynamic smoothing of the lot release in dependence of the fab loading situation. In this paper, we consider CONWIP, Workload Regulation, and CONLOAD, a rule developed for this study. For a recent survey on lot release rules, we suggest (Fowler, Hogg, and Mason 1998). The paper provides an outline of problems of controlling the performance of wafer fabs, presents a number of rules including advantages and problems, and offers a large body of literature references.

2.1 CONWIP

CONWIP regulation is a simple concept that is only based on counting the lots in production and capping this number. A lot is only allowed to enter the fab if the WIP level is lower than a given threshold Θ_{CONWIP}. Each time a lot enters the fab, Θ_{CONWIP} is increased by 1, and each time a lot leaves Θ_{CONWIP} is reduced by 1.

The positive effect on cycle times under this rule can be explained by Little's Law (Kleinrock 1975) which provides the general insight that average cycle times are directly proportional to average WIP. Thus, limiting WIP will limit cycle times.

Though being conceptionally simple and providing the basis of a number of success stories in improving fab performance, this method has some drawbacks. The threshold Θ_{CONWIP} is not a natural constant of the system. It has to be derived for each target system throughput or for each product mix individually. There are analytic approaches to find Θ_{CONWIP}, but the fine tuning has to be done by simulation or by analyzing a queueing network model of the fab. Each change of the set of machines may lead to a different threshold.

Another problem is the fact that WIP is a very coarse measure of the fab loading situation. The fab load for a WIP of 100 lots sitting in the queue of the first machine is fundamentally different to that for a WIP of 100 lots waiting in front of the last machine. Therefore, CONWIP works best for balanced fabs running already in steady state. For fabs ramping up and down products the control is less effective.

2.2 CONWORK

In this study, Workload Regulation is termed CONWORK (CONstant WORK). We apply CONWORK to improve the coarse picture of the fab loading situation. Here, we measure the amount of processing time at the bottleneck that is currently represented by the lots being processed in the fab. A lot is released to the fab if the current workload plus the total amount of bottleneck processing time of this lot is less than a given threshold Θ_{CONWORK}. As soon as it is released the workload is increased by the sum of bottleneck processing times of this lot. Each time a lot leaves the bottleneck workcenter the workload is decreased by its bottleneck processing time.

In comparison to CONWIP, the regulation is much finer because it is taken into consideration how much work a single lot will create for the bottleneck, i.e., the machine that limits the capacity of the fab. We obtain a better representation of the actual workload due to the update of the workload counter after leaving the bottleneck. But, similar to the CONWIP case, the threshold Θ_{CONWORK} is not a natural constant of the fab and has to be determined, for instance, by simulation. The rule can be extended to a multi-bottleneck scenario. With respect to a multi-product environment with product mix changes only little is known.

Although this rule provides a better picture of the loading situation of the fab than CONWIP, the rule does not reflect how the load is distributed over time. CONWORK does not distinguish between a lot that needs the bottleneck for 10 hours during a cycle time of 100 hours and a lot that offers a workload of 10 hours during 1000 hours although the second load produces only 10% of the fab load of the first.

2.3 CONLOAD

CONLOAD is a simple extension of CONWORK. Instead of considering the amount of work for the bottleneck workcenter, the amount of load for the bottleneck workcenter is computed, i.e., the sum of bottleneck processing times of the lot divided by the average cycle time of lots of this product. A new lot is allowed to enter the fab if the current bottleneck load plus the load introduced by the new lot is less than a given threshold Θ_{CONLOAD}. Each time a lot enters the fab, the bottleneck load is increased by the lot's load, and each time a lot leaves the fab it is decreased by the same amount. In contrast to the CONWIP and the CONWORK case, the threshold Θ_{CONLOAD} is a natural constant of the system. It is the target utilization of the bottleneck workcen-

ter times the number of bottleneck machines. For instance, if the maximum bottleneck load should be 95% and the bottleneck workcenter consists of 4 machines, then the threshold $\Theta_{\text{CONLOAD}} = 3.8$.

The only CONLOAD parameters that have to be determined in advance by simulation or queueing analysis are the average cycle times for each product. Compared to the other rules, however, we are able to work only with natural fab parameters or constants and not with artificial thresholds.

3 SIMULATION MODEL

Typical wafer fabs consist of several hundred machines producing tens of different products at a time. The wafers are manufactured according to recipes that contain several hundred processing steps. Due to the layered nature of semiconductors, the wafers visit sequences of machines several times, i.e., they are proceeding through the fab in cycles. Memory chips may have up to 30 layers. This cyclic visiting sequence of machines is responsible for a large part of the logistic problems of wafer fabs because lots of different cycles compete for the same machines.

To make a simulation study feasible with respect to running time, we require a fab model that shows the aforementioned behavior, but is considerably less complex in terms of the number of machines. Figure 1 shows the proposed factory model. It consists of a bottleneck workcenter, three delay units, and a control unit. The bottleneck workcenter determines the fab performance to a large extent (Atherton and Atherton 1995) and is therefore modeled in detail considering the number of machines, processing times, and dispatch rules. The rest of the machines are modeled as delay units. The control unit decides whether the required number of layers/cycles have been finished, and directs the lots to the fab exit or back to the second delay unit.

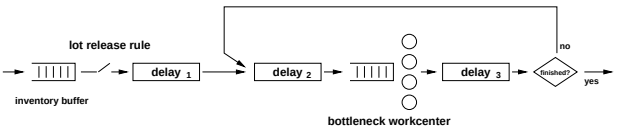


Figure 1: *Fab Model*

The bottleneck workcenter consists of four identical machines. The dispatch rule is FIFO. We did not choose due-date oriented dispatch rules in order to avoid problems in interpreting the results. Other studies (e.g., (Wein 1988)) indicate that it might be difficult to separate the effects of the lot release rule from those of the dispatch rule. The delay units are parameterized as follows where ERLA(k,m) denotes an Erlang-k distribution with mean m. All times are given in hours. The parameter ν is a scaling constant that depends on the product. It can be used to model an increase or decrease in the number of processing steps per layer.

Delay unit	Delay
$delay_1$	15+ERLA(5,$\nu \cdot 25$)
$delay_2$	20+ERLA(2,$\nu \cdot 30$)
$delay_3$	30+ERLA(3,$\nu \cdot 45$)

We consider the following 4 products where BNPT denotes the bottleneck processing time for one layer.

Product	BNPT	Layers	ν
1	1	9	1
2	1	7	1
3	1.5	9	1
4	1	9	1.3

We simulate three product mix change scenarios.

Decreasing number of layers. Start mix: 1,3,4. Final mix: 2,3,4.

Increasing bottleneck processing time. Start mix: 1,2,4. Final mix: 2,3,4.

Decreasing number of proc. steps/layer. Start mix: 2,3,4. Final mix: 1,2,3.

Each mix consists of three products where one product will be replaced by a new one. The products that are ramped up and down induce 30% of the target fab load. The two other products cause 35% of the target utilization each. For all experiments a target fab load of 95% applies. The lots are released to the inventory buffer uniformly, i.e. with constant interarrival periods.

Each simulation run lasts 6000 hours. At time 3000, the product mix change is introduced by stopping the release of the product to be replaced. The release of the new product starts instantaneously. For all considered scenarios it takes about 1300 hours until the last lot of the old product mix has left the fab.

Measurements of WIP, total number of lots, bottleneck queue length, and bottleneck utilization are taken from 2500 hours to 6000 hours. To reduce the amount of data and to facilitate the synchronization of the measurements from different replications, we apply the following method. The simulated time is divided into 10-hour intervals. For each 10-hour interval, we compute the time-based average of the above

performance measures, i.e., each value observed during this interval is weighted by the percentage of time during which it is kept. For each replication, we obtain a condensed sequence of 350 values ((6000-2500) hours/10 hours).

To obtain statistically useful results, each experiment is repeated 250 times. The curves shown in the rest of the paper are based on averaging the condensed sequences of 250 simulation replications. The 95% confidence intervals are reasonably narrow for this kind of transient measurements.

4 SIMULATION RESULTS

In this section we present the results of the comparison of the fab behavior during the product mix change scenarios for the CONWIP, CONWORK, and CONLOAD release rules. The reference model is a model without release control termed PUSH. The thresholds required for the CONWIP, CONWORK, and CONLOAD scenarios are set for the product mixes individually such that a fab load of 95% is guaranteed. The change of thresholds takes place immediately at time 3000 hours when the product mix starts to change.

Figure 2 shows the WIP and the total number of lots in the system where the new product has a lower number of layers. Figure 3 depicts the bottleneck utilization for this scenario under the regime of the four rules considered. The decrease in layers leads to a considerable increase in inventory and a temporal overload of the bottleneck workcenter in the PUSH case. CONWIP and CONLOAD reduce the amount of WIP and the bottleneck load at the cost of a longer transient phase until the fab reaches steady state for the new product mix. The PUSH fab is stable again at about 4500 hours, whereas the transient phase lasts at least until 6000 hours in the CONWIP and CONLOAD fabs. CONWORK shows the worst performance. It leads to a loss in capacity right after the release of new products to the fab starts. The WIP maximum is almost as large as in the PUSH case. The reason is that due to the decrease in layers the amount of workload per lot is also decreased. Thus, there is a tendency that more lots enter the fab than leaving it.

Figure 4 and Figure 5 show the performance measures when the bottleneck processing time of the new product is increased. For this scenario, there is no difference in behavior for the PUSH, CONWORK, and CONLOAD fab. We therefore omit the CONWORK and CONLOAD figures. Under CONWIP regime, however, the fab behaves worse than in the case without release control. The new product mix leads to a threshold decrease from 575 lots to 524 lots at time 3000 hours. Hence, lot release is throttled down for a considerable amount of time. As a consequence, bottleneck utilization goes down and throughput is lost. In turn, this leads to a long transient phase until stability is reached for the new product mix. The only way to solve this problem is to adapt the CONWIP threshold until all old lots have left the fab. The increase in control logic complexity, however, would be considerable.

In Figure 6 and Figure 7, the performance measures for the reduced steps per layer scenario are presented. In the PUSH fab, the WIP evolution shows no peculiarities. The bottleneck utilization is temporarily increased. In the CONWIP case, there is an increase in total number of lots at about 4000 hours. At this time, almost all old product lots have left the fab and the number of new lots leaving the fab is small due to the sharp decrease in WIP at time 3000 hours. Therefore, the total number increases since new lots are arriving uniformly at the inventory buffer. This phenomenon repeats itself about 1000 hours later with a smaller intensity. With respect to bottleneck utilization, CONWIP leads to a capacity loss that has to be compensated by a period of overload. At time 6000 hours, the fab is far from being running stable with the new product mix. For CONWORK, the fab shows a similar behavior as for CONWIP, but both the increase in WIP and the over-/underload situations are less intense. In contrast to CONWIP and CONWORK, the CONLOAD rule outperforms the PUSH rule. It improves the WIP situation and keeps the bottleneck utilization at the desired level.

With respect to the average and the variance of bottleneck queue length for the observation interval from 2500 hours to 6000 hours, the CONWIP rule improves the results of all other rules.

5 CONCLUSION AND OUTLOOK

In this paper, we presented CONLOAD, a new lot release rule for wafer fabs. This rule aims at keeping the bottleneck utilization at a given target level. The rule is conceptually simple and easy to implement.

We compare the fab performance under the regime of CONLOAD, CONWIP, or Workload Regulation (CONWORK), and without lot release control (PUSH) for three different scenarios. CONLOAD outperforms the other rules with respect to achieving the desired level of bottleneck utilization while inducing a smooth evolution of the WIP over time. This behavior also reduces the variations in cycle times and smoothes the lot departure process of the fab. As a

side result, we were able to show that CONWIP may lead to performance degradations for some product mix change scenarios.

There are a number of open issues with respect to the CONWORK rule. The rule requires the average cycle time of each product to work correctly. The sensitivity of the performance of the rule against wrong estimates for the average cycle times has to be assessed in a future study. The simple simulation model of this paper mimics typical characteristics of a real wafer fab. A typical wafer fab, however, is more complex and has more products. Therefore, CONLOAD should be implemented in a full fab model to check the performance in an environment that is closer to reality.

ACKNOWLEDGMENTS

The author would like to thank Frank Weissenseel for fruitful discussions and his valuable programming efforts.

REFERENCES

Atherton, L. F. and R. W. Atherton (1995). *Wafer Fabrication: Factory Performance and Analysis.* Boston: Kluwer.

Fowler, J. W., G. L. Hogg, and S. J. Mason (1998). Workload control in the semiconductor industry. Technical report, Arizona State University, IMSE, Tempe, AZ.

Hopp, W. J. and M. L. Spearman (1996). *Wafer Fabrication: Factory Performance and Analysis.* Chicago: Irwin.

Kleinrock, L. (1975). *Queueing Systems, Vol. 1: Theory.* New York: Wiley.

Lawton, J. W., A. Drake, R. Henderson, L. M. Wein, R. Whitney, and D. Zuanich (1990). Workload regulating wafer release in a GaAs fab facility. In *Proceedings of the International Semiconductor Manufacturing Science Symposium*, pp. 33–38.

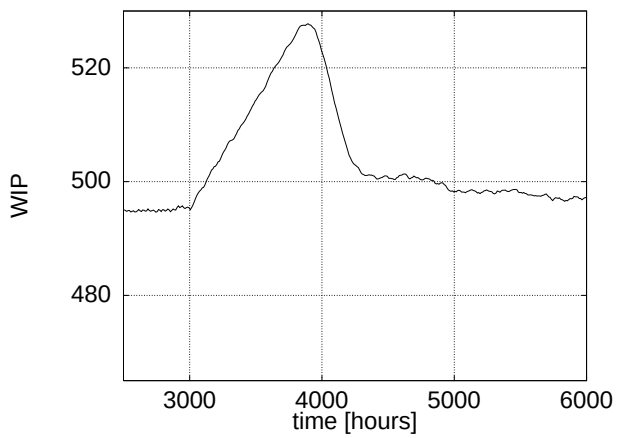
Rose, O. (1998). WIP evolution of a semiconductor factory after a bottleneck workcenter breakdown. In *Proceedings of the Winter Simulation Conference '98.*

Rose, O. (1999). Estimation of the cycle time distribution of a wafer fab by a simple simulation model. In *Proceedings of the SMOMS '99.*

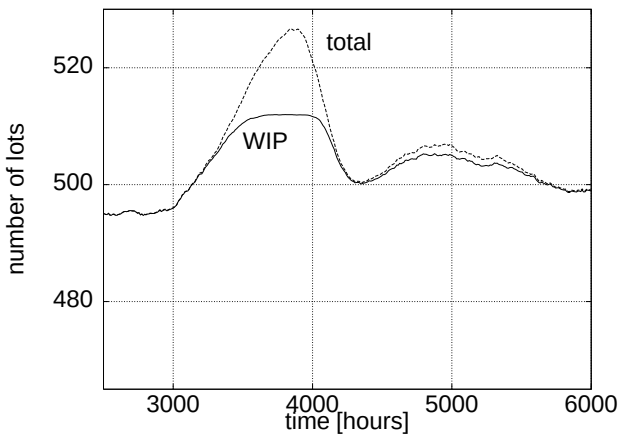
Wein, L. M. (1988). Scheduling semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing 1*(3), 115–130.

AUTHOR BIOGRAPHY

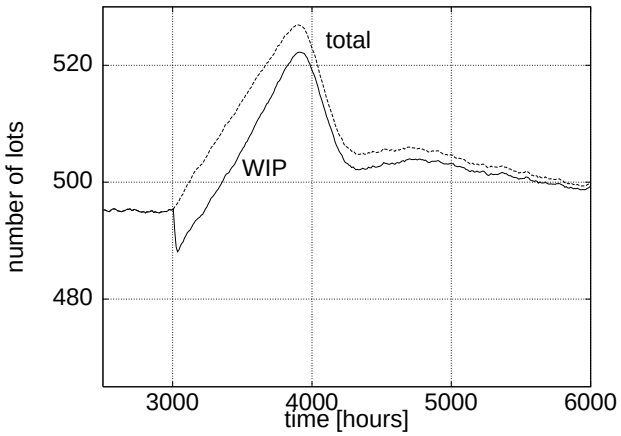
OLIVER ROSE is an assistant professor in the Department of Computer Science at the University of Würzburg, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from the same university. He has a strong background in the performance evaluation of high-speed communication networks. Currently, his research focuses on modeling and analysis of semiconductor and car manufacturing facilities. He is a member of IEEE, INFORMS, and SCS.



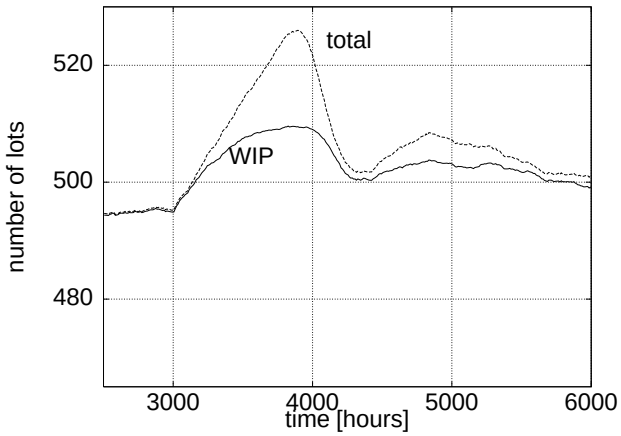
(a) PUSH



(b) CONWIP

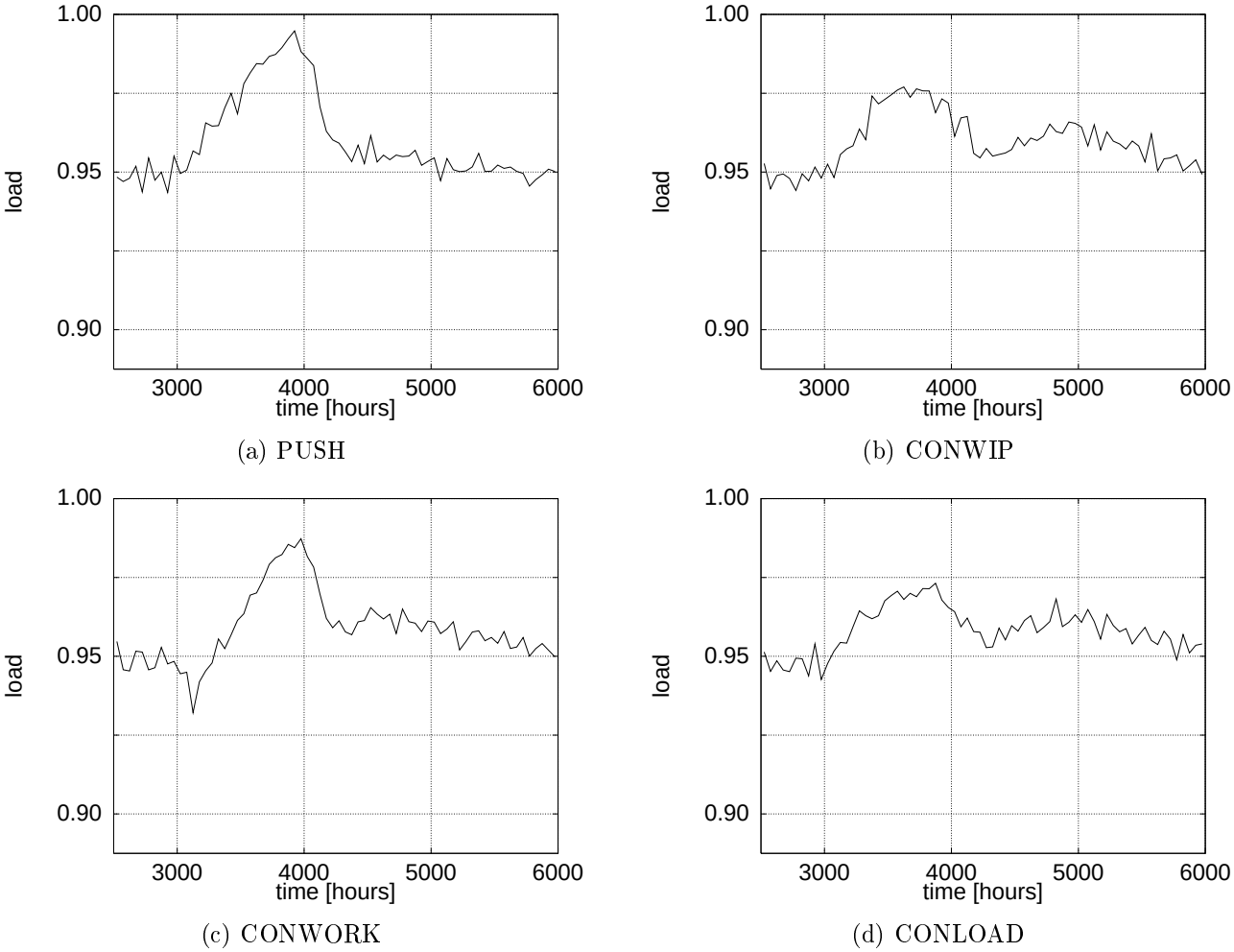


(c) CONWORK



(d) CONLOAD

Figure 2: WIP for Decreasing the Number of Layers



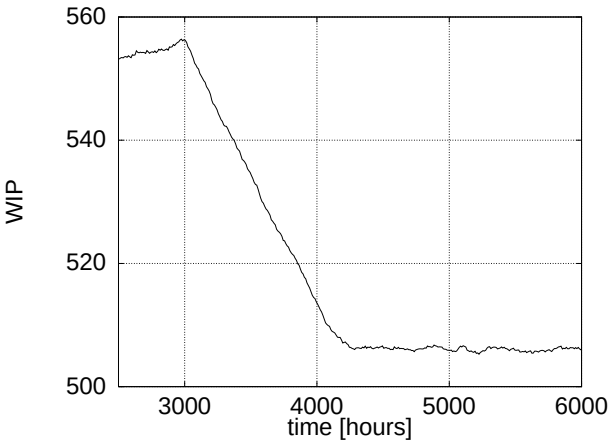
(a) PUSH

(b) CONWIP

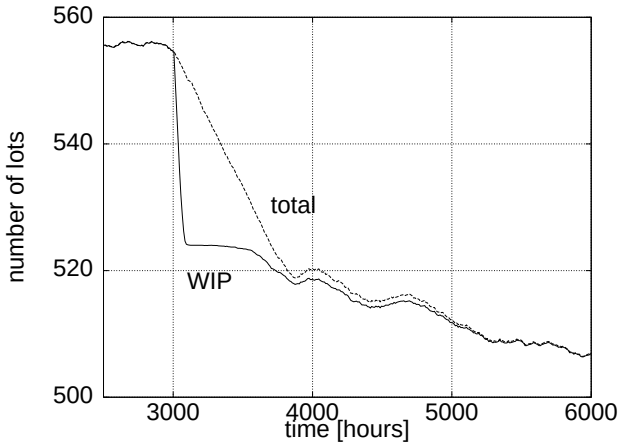
(c) CONWORK

(d) CONLOAD

Figure 3: Bottleneck Utilization for Decreasing the Number of Layers

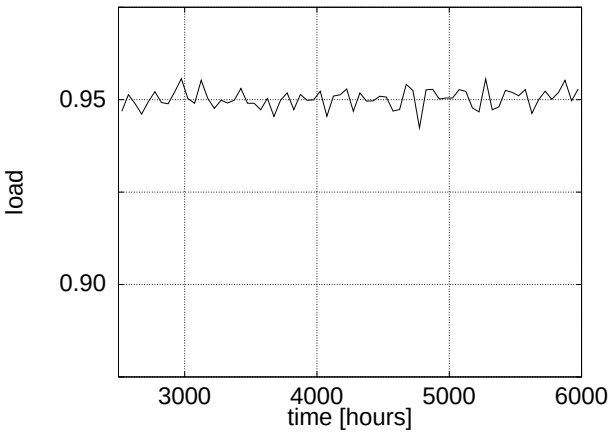


(a) PUSH

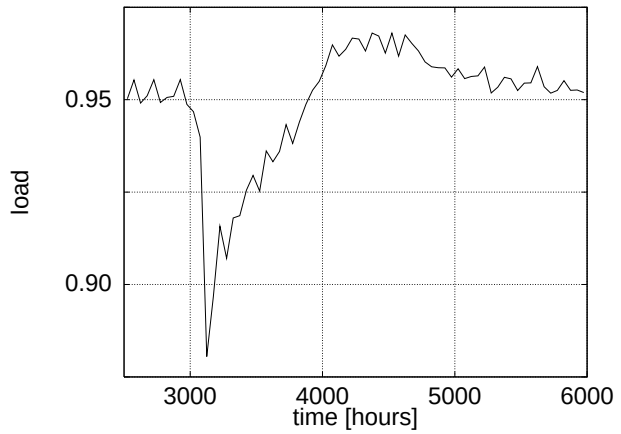


(b) CONWIP

Figure 4: WIP for Increasing the Bottleneck Processing Time

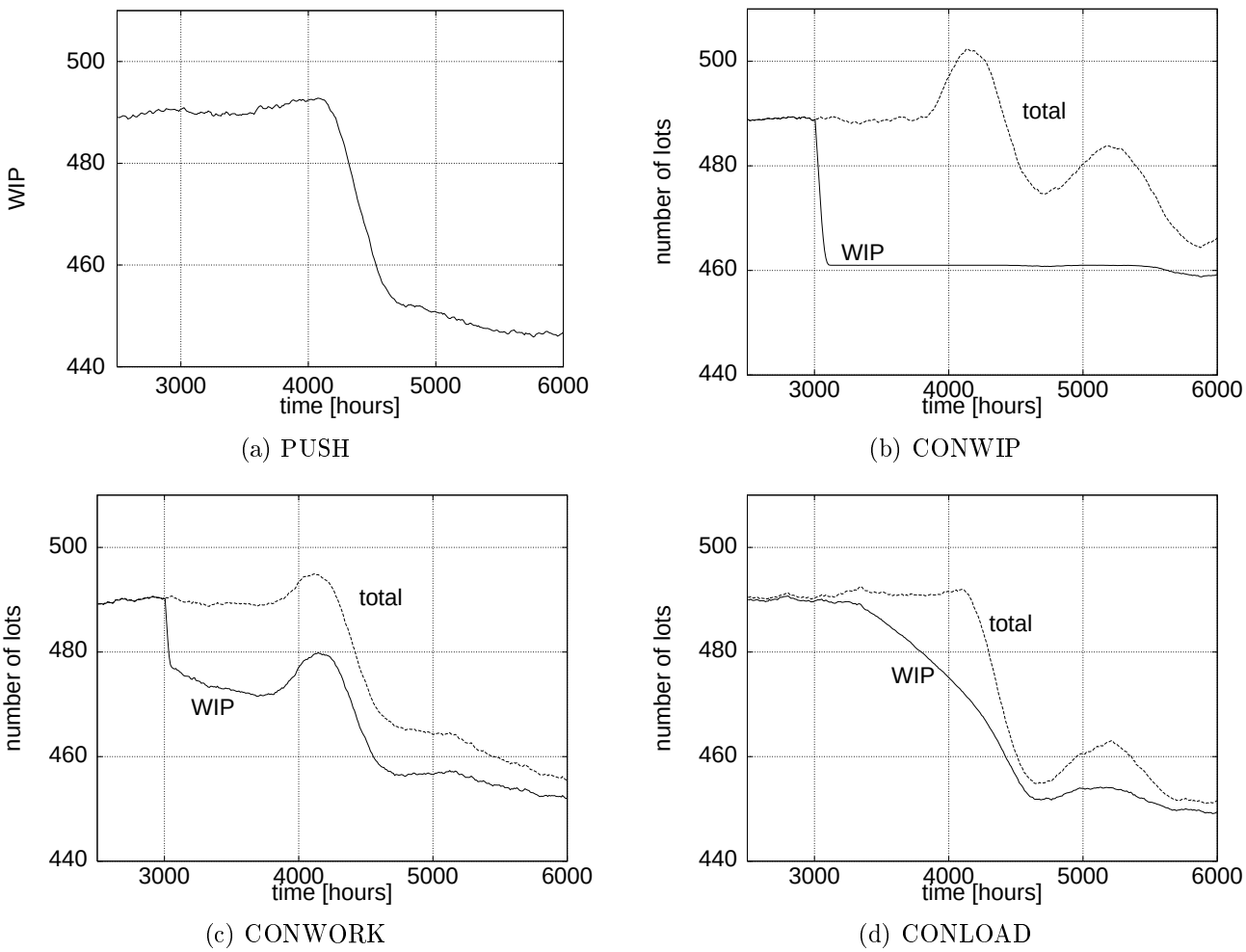


(a) PUSH



(b) CONWIP

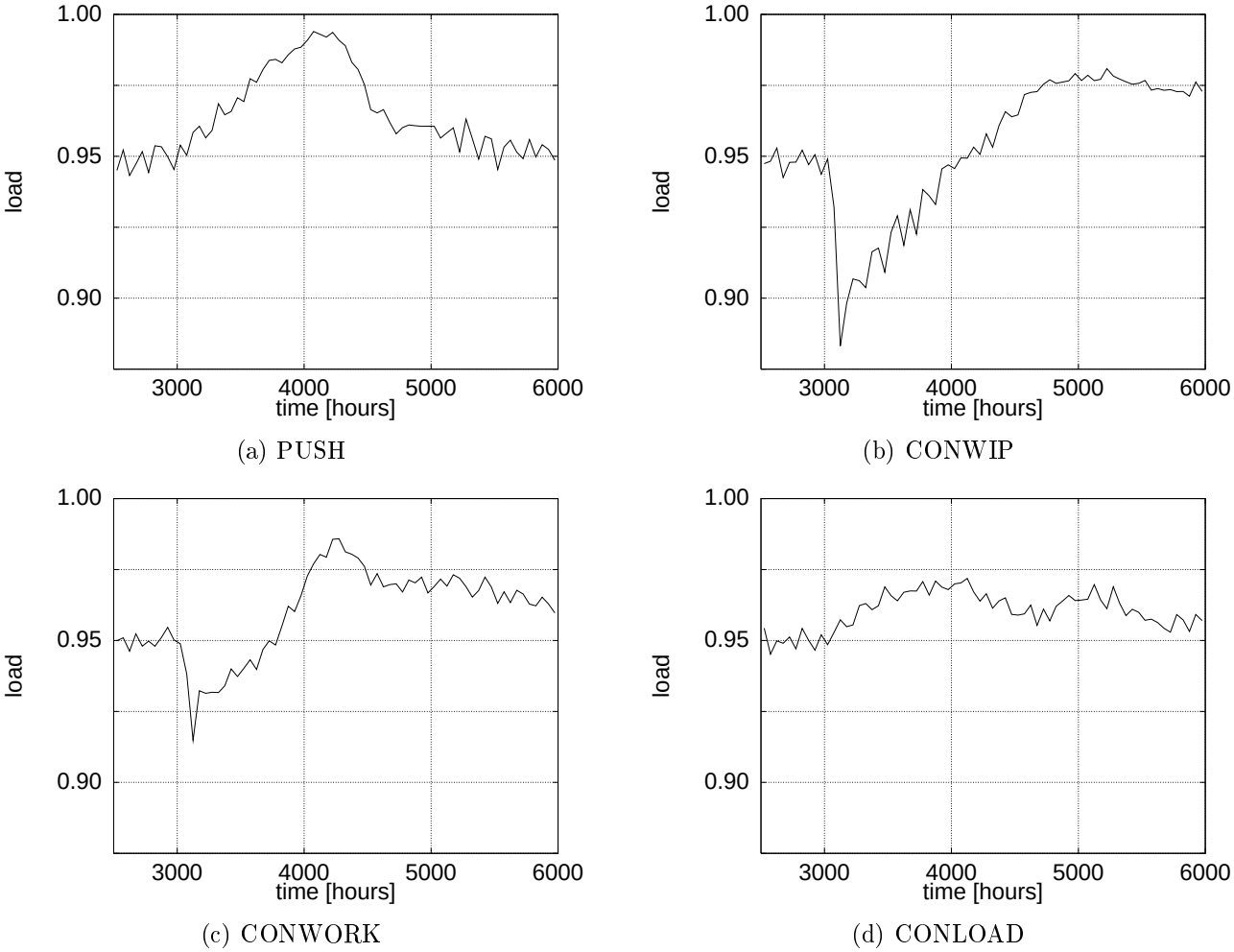
Figure 5: Bottleneck Utilization for Increasing the Bottleneck Processing Time



(a) PUSH (b) CONWIP

(c) CONWORK (d) CONLOAD

Figure 6: WIP for Decreasing the Number of Processing Steps Per Layer



(a) PUSH

(b) CONWIP

(c) CONWORK

(d) CONLOAD

Figure 7: Bottleneck Utilization for Decreasing the Number of Processing Steps Per Layer