Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

MANAGING WIP AND CYCLE TIME WITH THE HELP OF LOOP CONTROL

Steffen Kalisch
Robert Ringel
Jörg Weigang

AMD Saxony LLC & Co. KG
Wilschdorfer Landstr. 101
01109 Dresden, GERMANY

ABSTRACT

As an adaptation of the CONWIP concept, AMD has developed a heuristic approach to control the WIP in its wafer fabrication facilities (fabs). The so called "Loop Control" concept helps to utilize the installed equipment in an efficient way and reduces the overall cycle time. Two dynamic constraints (the WIP-Limit and the THP-Limit) are defined to limit the WIP (work in progress) per photo layer and to tackle high WIP situations at individual operations inside a photo layer. Loop Control has been evaluated with help of a fab simulation study prior to implementation in the fab dispatching system. Since its development three years ago, this system operates successfully in the AMD Dresden fabs.

1 INTRODUCTION

Dispatching the material in a semiconductor manufacturing line helps to produce material in time, to save overall cycle time and utilize installed equipment in an efficient way. These goals directly contribute to the lean manufacturing strategy that aims at the elimination of "Muda" (waste) in terms of overproduction, waiting and inventory. Years ago innovative dispatching approaches like Kanban or CONWIP have been successfully developed and introduced in other industries, especially in assembly lines. However, the complexity of wafer fabs seem to prohibit the direct application of these concepts.

In contrast to many other industries, wafer fabs process the material following a job-shop discipline with recurrent flows. This means every lot is processed repetitively on the same tools performing different operations. At each tool group, a dispatching task ranks the lots according to the goals mentioned above. Looking at the complete tool set of a wafer fab one could see hundreds of different machines in different configurations to be categorized in 10-15 characteristic tool types. Most of the tools are performing physical and chemical processes at the very state of the art of today's science and technology. These tools are running highly automated and closely monitored operations in order to detect exceptions as soon as possible. Typically fab tools fall out of the regular production state for maintenance or repair very often. This imposes an additional challenge on the dispatching system.

AMD has developed an adaptation of the CONWIP concept to be used in its Dresden fabs. A detailed analysis of the process flow characteristics enabled the design of a generic approach to control the WIP. This has been implemented in a fab simulation model for evaluation. Eventually AMD deployed Loop Control in the dispatching environment of the real fab. After a phase of closely monitoring, the system is running automatically for the last 3 years in Fab 30 (200 mm fab) and Fab 36 (300 mm fab). Loop Control has been implemented as part of the Shop Floor Control as suggested by Hopp and Spearman (1996).

2 THE LOOP CONTROL APPROACH

2.1 Basic Concept

CONWIP is a KANBAN-like production control system employed to limit the amount of material in the manufacturing line. It uses cards to control the number of pieces in the system. Each piece entering the system seizes a card – if all cards are taken, a newly entering piece has to wait until a previous pieces leaves the system releases its card. CONWIP controls the complete line using a single set of cards.

Modern wafer fabs are often highly loaded close to the capacity limit, because installed equipment is extremely expansive and must be used as much as possible. To use CONWIP in such an environment adaptations are required as indicated by Spearman et al. (1990). Similar indications are made by Marek (2001). Therefore several modifications to CONWIP are required in order to address the specifics of job-shop manufacturing systems like a wafer fab. Instead of imposing material limits to the complete manufacturing line, the system has to define these individual limits for designated sections along the line. This is an implication of the long cycle times spanning several weeks

for the complete process flow, as well as the frequent changes in the product mix and the continuous technology ramps.

An obvious segmentation of the semiconductor manufacturing process flow is given by the stack of photo layers. These layers can be used to define the flow sections that are to be controlled by particular material limits. Every photo layer starts with a mask operation performed by a stepper tool. These tools are the most expensive equipment in the fab and consequently often represent the fab bottleneck tools. Our proposed approach limits the amount of material inside a particular photo layer section of the flow by a number of pieces related to the production target for this photo layer. If a photo layer section accumulates enough material to exceed the target, the steppers are supposed process material in other photo layer sections instead of pushing more WIP into an already congested one. Depending on the current situation of the manufacturing line the production targets and the WIP inside the layers are not constant in time. Consequently, the limits per layer need to be adjusted over time. This concept is called WIP-Limit – it is active at the mask operations heading a photo layer section of the flow.

Often WIP piles occur at operations inside the layer. In such a case the WIP-Limit would eventually prevent material of being pushed into the layer. However, a WIP pile also eliminates the need of operations immediately feeding the pile to process material that would advance towards it. If the feeding operations are performed by shared tools, then it appears to be beneficial if these tools support other operations instead. Therefore, additionally to the WIP-Limit, the so called Throughput-Limit (THP) concept has been developed to address this situation. A THP-Limit is set at all operations inside the photo layers. They prevent the processing of material in operations that would stack onto a WIP pile in a subsequent operation.

Both concepts, the WIP-Limit and the THP-Limit, help to utilize the installed tools in an efficient way. They avoid pushing material into overloaded sections of the manufacturing line and instead pull material into other sections. This helps to balance the WIP along the line.

AMD has defined so called Loops in its manufacturing system. A Loop is a synonym for all operations between two mask operations including the starting mask operation. Loop Control is used to limit the amount of material entering the Loop with the help of the WIP-Limit. Yet it also tackles high WIP situations at operations inside a Loop using the THP-Limit.

Figure 1 shows the line status for the loop 31 – 33. Bars are used to indicate the production targets per operation. The box symbol line states the current WIP per operation. As shown in figure 1, the number of operations per loop is different for the individual loops. The current WIP at the operations shown in figure 1 is also very different.

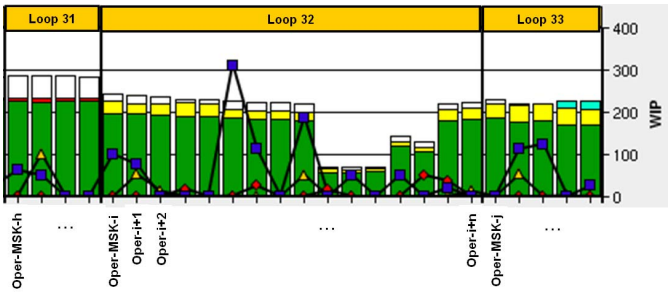

Figure 1: Sample Loops

A special challenge in the implementation of Loop Control has been to define the dispatching contexts for the Loops. For the AMD Dresden fabs the major product flows, covering 90% of the material, have been analyzed and the dispatching context was defined in terms of Process Flow – Route – Operation. So a Loop is specified on the basis of the dispatching contexts for all operations reaching from the mask operation of the layer up to the operation immediately proceeding the next photo mask operation.

2.2 The WIP-Limit

The Loop definition allows to determine the amount of material (WIP) inside the Loop. In order to set the WIP-Limit of a particular Loop the WIP for that Loop is compared to the Loop's production target (i.e., the requested number of wafers to process). Additionally, the cycle time of the Loop (i.e., the planned time for a lot to get processed by all Loop operations) has to be considered in the generation of the WIP-Limit per Loop, because the individual Loops differ in their number of operations (see Figure 1).
Definition of the WIP-Limit:

$$L_{WIP} = f \cdot T_{ShiftLoop} \cdot CT_{Loop}$$

f:	Factor to inflate the target value per shift to a value for a complete day and to control the sensibility of the limit (f= 2.0 … 2.3)
$T_{ShiftLoop}$ *:*	Average production target value per shift over all operations in the Loop
CT_{Loop} *:*	Accumulated cycle time of all operations in the Loop

The WIP-Limit is attached to all mask operations. An actual WIP level in a particular Loop above this limit ensures that there is enough material behind the mask operation to support the production target for this Loop. In this case the stepper should be used to process material going into an other Loop.

2.3 Throughput Limit

The THP-Limit has to be set up for all process operations. For that purpose the amount of material at every operation is compared to the production target for the operation. If the WIP is above the target ($T_{ShiftOper}$), the limit will be active. In this case shared tools at feeding operations are requested to process material for other operations and to avoid pushing more material to the stack. The current WIP and the production target as well as the plan cycle time per operation are used to calculate the so called "cycle time shadow" that is inflicted by the WIP pile. Material at all feeding operations within that shadow is not required to be processed at that point in time, since it would be pushed onto the stack. The corresponding shared tools are supposed to process other material in this case. As the pile decreases the shadow decreases as well and the feeding operations will resume processing at those shared tools.
Definition of the THP-Limit:

$$L_{THP} = T_{ShiftOper}$$

Definition the Cycle Time Shadow (by Littles Law):

$$CT_{Shd} = 12hrs \cdot \frac{WIP}{T_{ShiftOper}}$$

CT_{Shd} : This limit is only considered, if the WIP at the operation is above the production target per shift ($T_{ShiftOper}$). The CT_{Shd} value expresses in shift hours, how much material is waiting at the considered operation. The value is the ratio of WIP/ $T_{ShiftOper}$ multiplied by the duration of a shift.

WIP: Current amount of wafers at the considered operation

$T_{ShiftOper}$: The production target for the current shift at the considered operation

The example in Figure 2 shows a WIP of 1100 wafers at an Implant operation in Loop 11. The production target is assumed to be 350 wafers per shift. This results in a cycle time shadow of about 38 hours. According to the operation process time in this example the feeding operations Oper-i+2 and Oper-i+1 will be covered by the shadow. As processing at implant goes on, the WIP pile will lower, the shadow will shrink as well and material at the feeding operations Oper-i+1 and Oper-i+2 will be released.

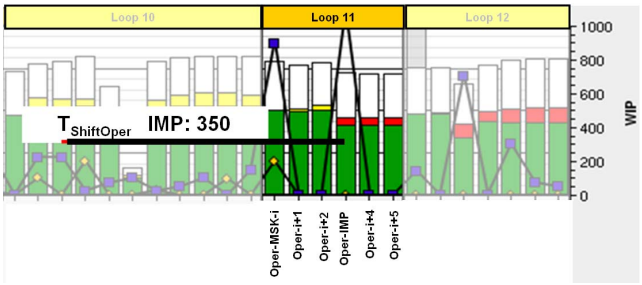

Figure 2: Example THP-Limit

2.4 Blocking or De-Prioritizing

The application of the WIP-Limit and the THP-Limit can be done in two ways. The first approach is to block material in case of an active limit. This means lots at the affected operations would not be processed as long as the limit is active even if the tools supporting these operations would starve due to this restriction. The second approach is to de-prioritize material at operations that are currently constraint by an active limit. In that case the affected material would be ranked down to the bottom of the corresponding dispatch lists. However a shared tool running out of other material would resume processing the lots at these currently constraint operations. Similar considerations are made by Rust (2007).

3 SIMULATION STUDIES

3.1 Model characteristics

AMD uses the Applied Materials™ ASAP™ simulation tool. ASAP is employed to model the capacity aspects of the fab. Therefore the tool set descriptions, the process flows, the current WIP, the lot starts and the tool availability data as well as reticle sets and dedication information feeds the generation of the simulation model. ASAP's customization framework software interface allows the experienced user to code model extensions. The AMD custom extensions enhance the standard tool models in order to reflect AMD's specific fab tool characteristics as well as reticle dispatching, implanter species management, tool dedications and others. AMD specific dispatching rules based on targets and due dates are attached to ASAP by custom extensions as well. On a daily bases model input data is automatically retrieved from the fab data repositories. Major input files are checked and monitored automatically, too.

In order to evaluate Loop Control the simulation framework required additional input files providing data of the Loop-based dispatching contexts. The custom extensions have been enabled to perform cyclic checks of WIP-Limit and THP-Limit as described below. In case of active limits the customized dispatching rules have been enabled to block the material at the affected dispatching contexts.

3.2 Proof of concept

On a high level of detail a 7 days full fab simulation run takes about 15 minutes. It was decided to perform the evaluation study with a pre-loaded fab model and to run for a 7 day period. The current WIP and the lot starts as well as the tool set of the selected work weeks have been used in the model. Another approach would have been to start the simulation with an empty fab, fill it up with WIP until a steady state is reached and then start recording the simula-

tion outputs. That second approach was disregarded, since it would have required more simulation run time and would have generated a WIP distribution that differs from the real fab snap shot situation.

WIP-Limit Evaluation: During the experiment, a front-end implantation layer shows a high WIP situation lasting two days. The production target value is 375 wafers and the cycle time for this layer is 1.2 days in the example. The resulting WIP-Limit value is 900 wafers (using f=2). As seen in Figure 3, the WIP per layer is limited by the WIP-Limit control to a value of 900 wafers, whereas the system without Loop Controls yields a WIP in the layer of above 1000 wafers per day. In the Loop Control case the upstream steppers do not process material heading towards the implant layer until the amount of material in this layer drops below the WIP-Limit value.

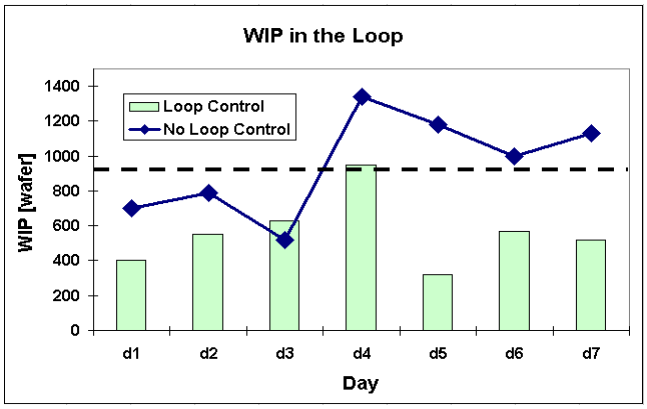

Figure 3: The effect of the WIP-Limit

At day 4 the WIP-Limit is active for 12 hours and at day 5 it is active for 8 hours in the above example. As shown in figure 4 the number of processed wafers at mask operations is much less at these days, running Loop Control.

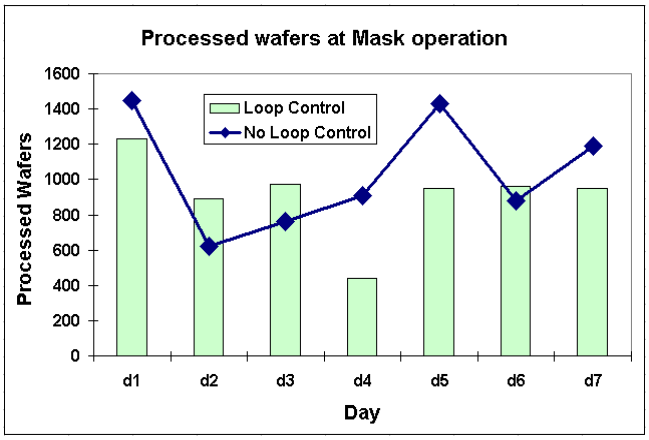

Figure 4: An active WIP-Limit effects the processed wafers at the mask operation

Figure 5 shows the active phases of the WIP-Limit within the period of 7 days. The limit value in this example is 840 wafers.

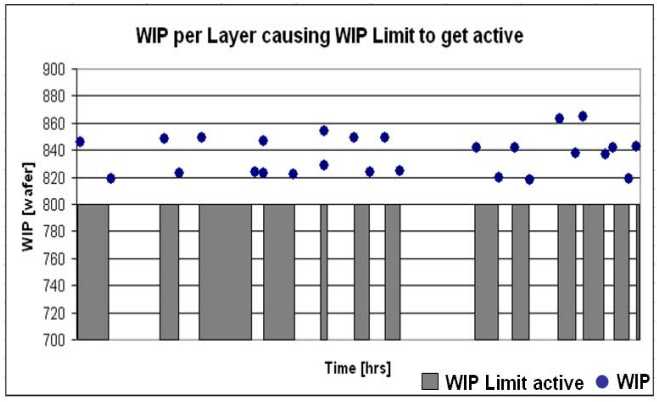

Figure 5: Activation and Deactivation of the WIP-Limit

THP-Limit Evaluation: During the experiment a special tool down situation occurs in an etch tool family for about four days. The availability of this tool set drops below 40% causing the WIP for the supported etch operation to pile up to over 2000 wafer in the non-Loop-Control case. In the process flow the etch operation follows immediately to the mask operation. Consequently the THP-Limit of the affected etch operation gets active, because the limit value is at 400 wafers.

Figure 6 shows the break down of processed wafers at etch at day 4 caused by the low tool availability in this station family. This happens in both scenarios, Loop-Control and No-Loop-Control.

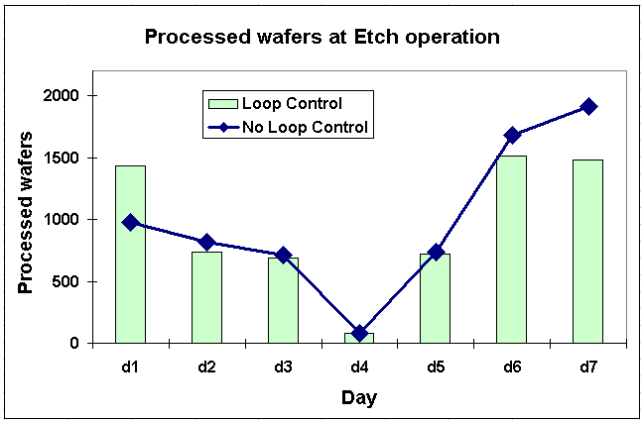

Figure 6: Low availability at Etch operation

For the case of activated Loop Control Figure 7 illustrates how the WIP at the etch operation is stabilized, because material is waiting at the mask operation. The stepper tools supporting this mask are not idle while THP-Limit is active. They process material at other operations in the meanwhile, as shown in Figure 8.

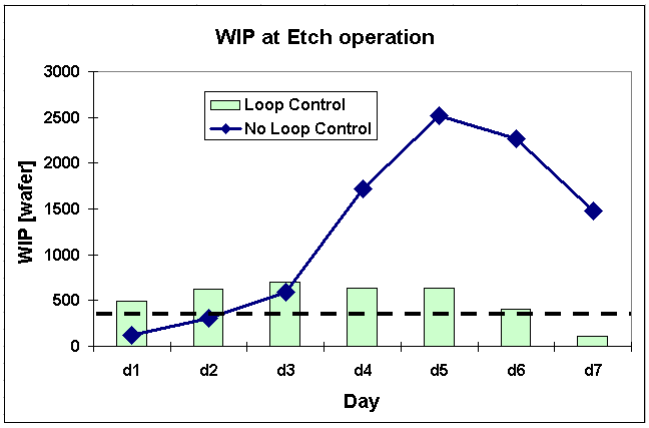

Figure 7: Active THP-Limit levels the WIP per operation

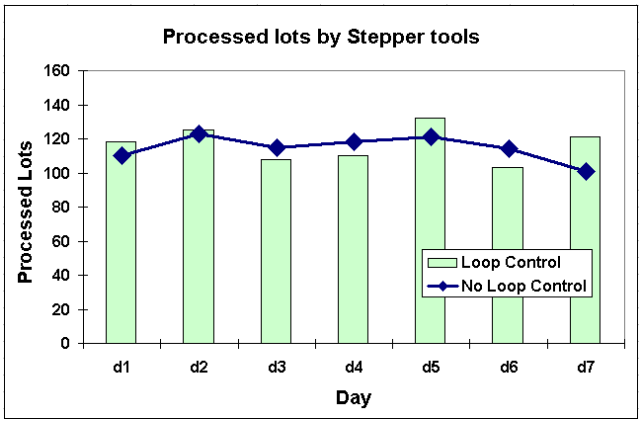

Figure 8: Processed lots by stepper tools

3.3 Evaluation

The goal of the simulation study is to check whether Loop Control helps to increase the number of processed lots and reduces the cycle time. Furthermore, it should be checked whether Blocking material is better then De-Prioritizing or not.

Several work week scenarios have been simulated using Loop Control with different parameters and have been compared to the simulation runs not using Loop Control. Cycle time and processed lots in both cases have been considered in order to evaluate Loop Control. Additionally, several artificial long tool down scenarios have be created in the simulation study in order to see how Loop Control manages these exceptions.

Figure 9 shows that blocking material in case of active limits leads to a reduced number of processed lots and to an increased cycle time compared to de-prioritizing material. De-Prioritizing material means, to rank down lots in the dispatch list of the contexts for which the WIP-Limit or the THP-Limit is currently active. So in case a shared tool has only material at currently limited contexts, it would process that material in case of de-prioritization whereas stay idle in the blocking case.

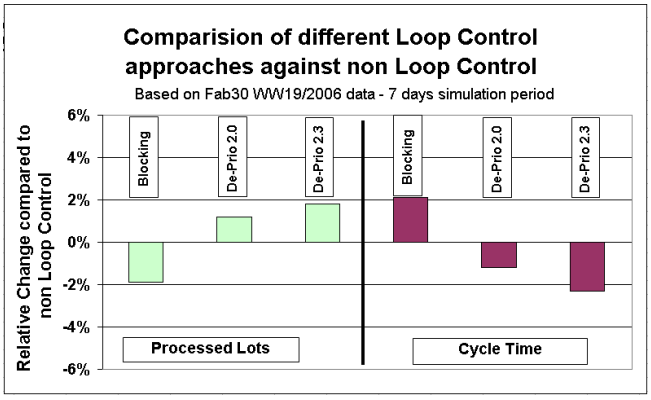

Figure 9:Processed lots and cycle time for different Loop Control approaches

For de-prioritization the factor f is used to control the sensibility of the WIP-Limit. A value f=2.0 implies an activation of the limit at a lower WIP level than a value f=2.3. Especially in terms cycle time the results in Figure 10 shows the best results for value f=2.0. Activating the limits at lower levels can cause more balanced WIP situations and helps to avoid higher WIP piles. This results in lower cycle time.

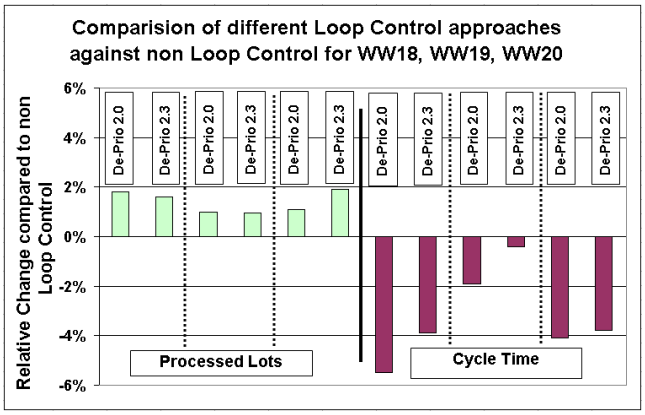

Figure 10: Loop Control simulation results for input data taken from different work weeks

Finally, line exceptions caused by tool down situations of varying length in different sections of the manufacturing line have been simulated. Also in these scenarios Loop Control helps to reduce the cycle time as shown in Figure 11.

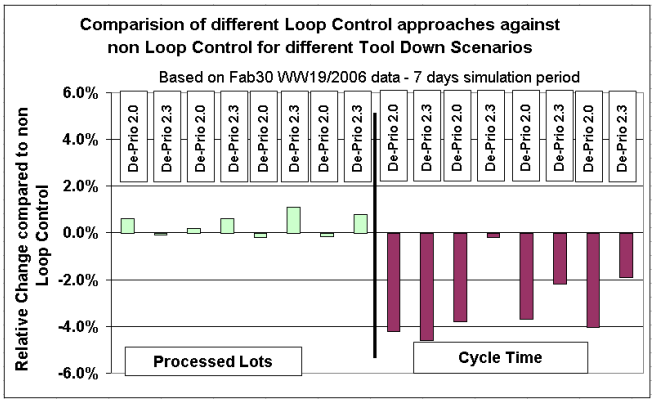

Figure 11: Loop Control simulation results for long tool down scenarios

The overall simulation results indicate that Loop Control with blocking of material at mask operations results in more cycle time and less activities compared to both, Loop Control without blocking and operating without Loop Control. De-Prioritizing materials helps to process more lots and results in a reduced cycle time.
After analyzing the simulation results, it was decided to deploy Loop Control to production using the factor f=2.0.

The expectation was to improve fab activities and cycle time by about 3% when Loop Control gets installed.

4 DEPLOYMENT IN THE FAB

4.1 Deployment Steps

In order to set up Loop Control in a real fab the following actions and calculations are required:

1. Update the daily production target for the next day and determine the WIP-Limit and THP-Limit values (automatically every day at 7pm)
2. Check these limits and set or re-set them – Quote those limits in the corresponding production control reports (automatically every hour)
3. Monitor the limits and deactivate them in case of problems (manual)

Loop Control was installed in the fab dispatching environment in following steps:

1. Implementation of the WIP-Limit
2. Quotation of active WIP-Limits in a special web-based report for observation and deactivation by production control staff in case of problems
3. Implementation of the THP-Limit
4. Quotation of all active THP- and WIP-Limits as in 2)
5. Fully automated execution of Loop Control – no manual deactivation of individual limits

These steps were taken in a time frame of about 4 weeks while Fab 30 was fully loaded.
During this phase Loop Control was closely monitored by the dispatching and production control staff. There were no major problems applying Loop Control.

After a short period of observation the new approach was fully accepted by production control as well. Particularly the similarities to the formerly used "Key Opers" (prioritize material) and "Limit Opers" (de-prioritize material) helped to understand and accept the Loop Control concept as a consistent way to exercise the same control automatically. As a result manual interaction at this point was no longer necessary.

For obvious reasons, it is difficult to measure the contribution of a particular material control strategy to the fab performance in terms of activities and cycle time:

- The fab is not a stable system – Constantly changing characteristics like product mix, tool set, technology and manufacturing flows prohibit stable evaluation conditions.
- Disabling an efficient dispatching strategy for the purpose of measuring its performance impact is not reasonable.

Nevertheless, the impact of Loop Control to the fab cycle time was estimated by the fab management as a reduction of about five percent. However the best indication for the success of the presented concept is, that Fab 36 management requested the installation of Loop Control during a phase of increasing fab load in order to help balancing the line.

5 SUMMARY

The simple idea behind Loop Control is to avoid pushing material into a (temporary) dead end (WIP piles) of the manufacturing line. For that purpose, a comparison of the current WIP situation and the production targets identifies those "dead ends". Dispatching helps to use the shared tools to process material in non-congested process flow sections until the WIP piles are lowered.

Therefore Loop Control can be seen as a special adaptation of the CONWIP concept for the application in a semiconductor manufacturing line. A special feature of Loop Control is that the defined limits for WIP and throughput are not constant over time. They are getting updated daily according to the production targets. Critical for the success of the Loop Control implementation is a detailed analysis of manufacturing flows. This analysis yields the dispatching contexts for which those limits apply to.

Another key feature is the alignment of Loop Control with the production target in terms of wafers to be processed.

The presented approach is successfully running in AMD's Dresden Fabs, because it runs automatically and

the idea behind is completely understood by the people there.

REFERENCES

Hopp, W., and M. Spearman. 1996. *Factory Physics - foundations of manufacturing management.* The McGraw-Hill Companies, Inc.

Marek, R. P., D.A. Elkins, and D.R. Smith. 2001. Understanding the fundamentals of Kanban and CONWIP pull systems using simulation. In *Proceedings of the 2001 Winter Simulation Conference*, 921-929.

Rust, K. 2007. Next generation Fab modeling using lean concepts. *Applied Materials Worldwide Software Symposium*, San Diego.

Spearman, M., D. Woodruff, and W. Hopp. 1990. CONWIP: a pull alternative to Kanban. *International Journal of Production Research* 28: 879-894.

AUTHOR BIOGRAPHIES

STEFFEN KALISCH received a M.S. in Industrial Engineering from Dresden University of Technology, Germany. Since April 2003 he is working with AMD Saxony LLC & Co. KG in Dresden. He is responsible for WIP management and simulation. In this matter he has enhanced the dispatching system using Brooks Real Time Dispatcher. <steffen.kalisch@amd.com>

ROBERT RINGEL is an Industrial Engineer at AMD Saxony LLC & Co. KG in Dresden, where he is responsible for the Fab simulation models. In 1996 he received a degree in Computer Science at the University of Technology Dresden, Germany. Before AMD he worked at the university doing research in computer simulation for traffic technology. Later he became a software developer for cellular radio in Berlin. <robert.ringel@amd.com>

JÖRG WEIGANG received a M.S in Computer Science from Dresden University of Technology, Germany. Since November 1997 he is working with AMD Saxony LLC & Co. KG in Dresden. In his current position as Member Technical Staff he is responsible for WIP management and simulation. In this matter he has mainly developed the dispatching system of AMD's Fab 30 and Fab 36 using Brooks Real Time Dispatcher. <joerg.weigang@amd.com>