Proceeding of the 2004 American Control Conference
Boston, Massachusetts June 30 - July 2, 2004

A Model Predictive Control Strategy for Supply Chain Management in Semiconductor Manufacturing under Uncertainty

Wenlin Wang*, Daniel. E. Rivera*[1], Karl G. Kempf† and Kirk D. Smith‡

*Department of Chemical and Materials Engineering
Control Systems Engineering Laboratory,
Arizona State University, Tempe, Arizona 85287-6006

†Decision Technologies,
‡Components Automation Systems,
Intel Corporation,
5000 W. Chandler Blvd., Chandler, AZ, USA 85226

Abstract

Model Predictive Control (MPC) is presented as a tactical decision module for supply chain management in semiconductor manufacturing. A representative problem which includes distinguishing features of semiconductor manufacturing supply chains, such as material reconfiguration and stochastic product splits, is examined. Fluid analogies are used to model the supply chain dynamics, with stochasticity and nonlinearity occurring on the throughput time, yield and customer demand. Given inventory targets and capacity limits, MPC using linear time invariant models can make the system outputs track the targets and improve customer service levels. The flexibility provided by the choice of tuning parameters in MPC to achieve better performance and robustness in semiconductor manufacturing supply chain management is demonstrated.

1 Introduction

The global market in the 21st century is electronically connected and dynamic in nature. Therefore, companies are trying to improve their agility level with the objective of being flexible and responsive to meet the changing market requirements [1]. Supply chain management (SCM) is an important consideration in today's manufacturing industries because of the vital role it plays in distributing resources and generating profits. The paper by Kempf [2] describes the role of integrated decision policies for improving supply chain management in the semiconductor industries.

Recent work of Braun *et al.* [3] using Model Predictive Control has shown it as an attractive method for inventory control in supply chains. The work shows the effectiveness of a partially decentralized MPC structure under model mismatch and demand forecast error in a deterministic environment. In Wang *et al.* [4, 5], a centralized MPC strategy is successfully used in semiconductor manufacturing SCM to track inventory targets and satisfy customer demands. The appeal of MPC for SCM can be summarized as follows: as an optimizer, MPC can minimize or maximize an objective function that represents a suitable measure for supply chain performance. As a controller, MPC can be tuned to achieve stability, robustness, and performance in the presence of plant/model mismatch, disturbance and uncertainty which affect the system. This work focuses on applying a centralized MPC strategy to a representative problem described in [2] that incorporates nonlinearity and short timescale stochasticity in the process.

The paper is organized as follows. In Section 2, the Model Predictive Control formulation and its application to supply chains are discussed. In Section 3, a typical problem involving distinctive features of semiconductor manufacturing is studied. Simulation results showing proof of concept for the MPC approach are discussed. This paper concludes with a discussion of the flexibility and advantages of using MPC in SCM.

[1]To whom all correspondence should be addressed. phone: (480) 965-9476 fax: (480) 965-0037; e-mail: daniel.rivera@asu.edu

0-7803-8335-4/04/$17.00 ©2004 AACC

2 Model Predictive Control

Model Predictive Control is an optimization-based control scheme. Its formulation integrates optimal control, stochastic control, control of processes with dead time and multivariable control. One of the advantages of using MPC is that it can easily handle constraints on both manipulated and control variables. The MPC controllers considered in this paper rely on a linear state-space model. The manipulated variables $u(k)$ are the starts for the manufacturing nodes. The customer demands are considered as the measured disturbances with anticipated forecast. The controlled variables are the inventory levels $y(k)$ which have reference levels $r(k)$ and the Work-In-Progress (WIP) in each manufacturing node, which have high and low limits only (without setpoints).

As a receding horizon algorithm, at each time instant t, the controller considers the previous information on inventory levels, actual customer demands, starts and future information on inventory targets, forecasted customer demand to calculate a sequence of future starts by solving the following optimization problem.

$$\min_{\Delta u(k|k)\ldots\Delta u(k+m-1|k)} J \tag{1}$$

where the individual terms of J correspond to:

$$\begin{aligned} J &= \overbrace{\sum_{\ell=1}^{p} Q_e(\ell)(\hat{y}(k+\ell|k) - r(k+\ell))^2}^{\text{Keep Inventories at Inventory Planning Setpoints}} \\ &+ \overbrace{\sum_{\ell=1}^{m} Q_{\Delta u}(\ell)(\Delta u(k+\ell-1|k))^2}^{\text{Penalize Changes in Starts}} \\ &+ \overbrace{\sum_{\ell=1}^{m} Q_u(\ell)(u(k+\ell-1|k) - u_{target}(k+\ell-1|k))^2}^{\text{Maintain Starts at Strategic Planning Targets}} \end{aligned} \tag{2}$$

subject to the constraints on the starts, the change rate of starts , the inventory levels and WIPs. Here p is the prediction horizon and m is the control horizon. $Q_u, Q_{\Delta u}, Q_e$ are penalty weights on the control signal, move size and control error, respectively. This problem can be solved by standard quadratic program algorithms.

Following the receding horizon principle, MPC applies the first element of the calculated control action to the system. After new measurements are available, a new optimization problem is solved. The use of future setpoint and disturbance changes in MPC is referred to as anticipative action when the estimated values of these signals are known in advance. Making use of anticipation in the controller is a significant contributor to improved performance. With proper tuning, an MPC approach relying on linear models with fixed throughput time and yield can achieve good performance in spite of the very nonlinear and stochastic characteristics in manufacturing process. This will be shown in the following sections.

3 The Assembly/Test2 stochastic split problem

A representative semiconductor manufacturing supply chain for a problem involving the manufacture and demand of two products with different speeds is considered. The fluid analogy corresponding to this problem is shown in Figure 1. It contains one Fab/Test1 node $M10$, one Assembly/Test2 node $M20$, two Finish/Pack nodes $M30$, one Assembly-Die Inventory (ADI) $I10$, one Semi-Finished Goods Inventory (SFGI) for high speed devices $I20$, one SFGI for low speed devices $I21$, one Components Warehouse (CW) for high speed devices $I30$ and one CW $I31$ for low speed devices. Two $M40$ nodes represent shipments with one day delay and no uncertainty. The valves from $C35$ to $C39$ are the starts for factories. $C40$ and $C41$ are where customer demands enter. Items coming into $M20$ through $C36$ will be split into two bins, one is made up of high speed components, while the other one has low speed devices. The number of items in each bin is determined by a split factor in A/T2 which is stochastic and can have different average and variance values. Besides meeting the fast device demand D, fast devices in $I20$ can also be used to make slow devices to meet the demand E. Devices in $I21$ can only be used to meet the customer demand E. In other words, if there are more than enough products available to meet customer demand D while not enough to meet E, through $C38$ some fast devices will be transferred from $I20$ to $I31$ to meet the demand E. The opposite direction is not allowed. Excess devices in $I21$ will be discarded through $C90$ if the inventory level reaches a maximum.

The controlled variables are the inventories, $I10$, $I20$, $I21$, $I30$ and $I31$. The associated variables are Work-In-Progress (WIP) for $M10$, $M20$ and $M30$. The manipulated variables are the starts for all the factories, $C35$, $C36$, $C37$, $C38$ and $C39$. Customer demands for high speed devices D and for low speed devices E are treated as measured disturbances coming into system through $C40$ and $C41$ respectively. Although we do not know the exact customer demand in future, a reasonable forecast can be achieved. This forecast is used

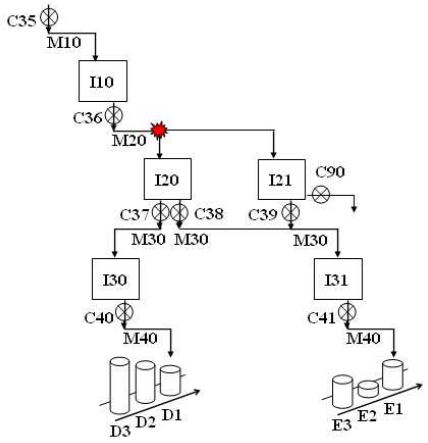


Figure 1: The Assembly/Test2 stochastic split problem

as an anticipation for future measured disturbance in MPC. There are errors between the actual demand and the forecast. However, the approximate anticipation can still improve the system performance. The nominal controller model used for inventories in this problem is based on a material balance. The representative equations for $I20$ and $I21$ which are related to the split factor can be written as follows

$$\begin{aligned} I_{20}(k+1) &= I_{20}(k) + Y_2 C_{36}(k-\theta_2)\cdot\alpha \\ &- C_{37}(k) - C_{38}(k) \\ I_{21}(k+1) &= I_{21}(k) + Y_2 C_{36}(k-\theta_2)\cdot(1-\alpha) \\ &- C_{39}(k) \end{aligned} \quad (3)$$

Here Y_2 stands for the yield of $M20$, θ_2 is the average throughput time in $M20$ and α is the split factor which is assumed in simulation as a random number with a uniform distribution.

The representative model for the WIP of the manufacturing node $M30$ for low speed devices can be described in the following equation.

$$\begin{aligned} WIP_{30}^{low}(k+1) &= WIP_{30}^{low}(k) + C_{38}(k) + C_{39}(k) \\ &- C_{38}(k-\theta_3) - C_{39}(k-\theta_3) \end{aligned} \quad (4)$$

where θ_3 is the throughput time for $M30$. In this formulation, the nominal model for the controller is linear in nature with fixed throughput time and yield. In the simulation model, the throughput time of $M10$ is nonlinearly dependent on the load or WIP as shown in Figure 2. It varies uniformly between three different load ranges (30 to 32 days at 0 to 70% load, 32 to 38 days at 70 to 90% load, and 35 to 45 days at 90 to 100 % load). The stochasticity can be demonstrated by the response of outflow and WIP of $M10$ to changes in starts as shown in Figure 3. One can find the stochasticity on the outflow of the F/T1 is dependent on the input magnitude which determines the load in the factory. For the other factories, the throughput time is only a uniformly distributed number varying from 5 to 7 days for Assembly/Test2 node and 1 to 3 days for Finish node. Yield rates also vary uniformly distributed from 0.93 to 0.97, 0.98 to 0.99 and 0.985 to 0.995 for $M10$, $M20$ and $M30$ respectively. The inventory targets are 3306, 1102, 551, 351 and 176 units for $I10$, $I20$, $I21$, $I30$ and $I31$ respectively.

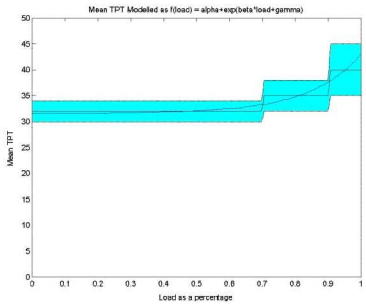


Figure 2: Nonlinear relationship between load and throughput time

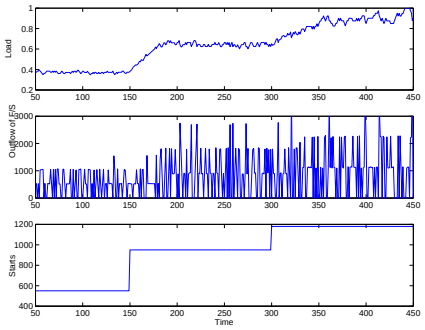


Figure 3: The Fab/Test1 node response to various step changes in starts

The constraints are enforced to keep high and low limits on inventory levels, WIPs, starts and the change of starts.

$$I_*^{min} \le I_*(k+i|k) \le I_*^{max} \quad (5)$$

$$WIP_*^{min} \le WIP_*(k+i|k) \le WIP_*^{max} \quad (6)$$

$$i = 1, 2, \ldots, p$$

$$0 \le C_*(k+j|k) \le C_*^{max} \quad (7)$$

$$\Delta C_*^{min} \le \Delta C_*(k+j|k) \le \Delta C_*^{max} \quad (8)$$

$$j = 1, 2, \ldots, m \quad (9)$$

$*$ represents the index for different inventories, WIPs and starts. The constraints are forced over the time

horizon. Here the prediction horizon p is 70 days and control horizon m is 60 days.

The two customer demands for fast and slow devices are stochastic with different means but similar variances. Usually the demand for the slow device is higher than that for the fast device; the average of the slow device demand is larger than that of the fast device demand, as shown in Figure 4. Depending on

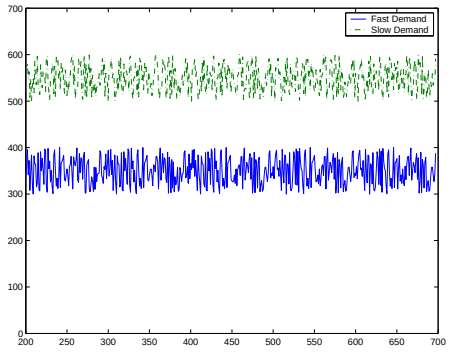


Figure 4: Demand sets in Assembly/Test2 stochastic split problem

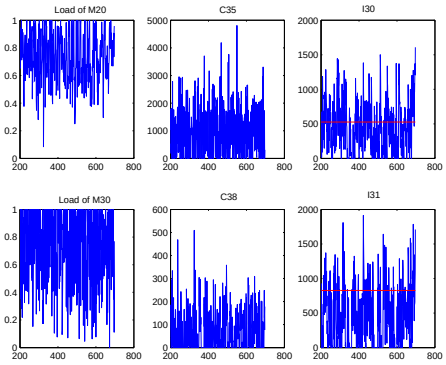
the split in Assembly/Test2, we will have different amounts of products to meet different customer demands. Based on the split factor in the A/T2 node, we have developed three cases for study. The average of high speed demand to that of low speed device is 0.39/0.61. In Case 1, the split is balanced to the demands, 0.39/0.61. Because of space considerations, we do not discuss this case here; refer to [5] for details. In Case 2, the average split is 0.49/0.51. More high speed devices are generated in this case and they have to be reconfigured to meet low speed device demand. In Case 3, the average split is 0.29/0.71 which implies not enough high speed devices are generated. The total amount of material processed in $M10$ has to increase to meet the high speed device demand. In all three cases, the variance of the split is 0.1.

3.1 Case 2 (Reconfigure High)

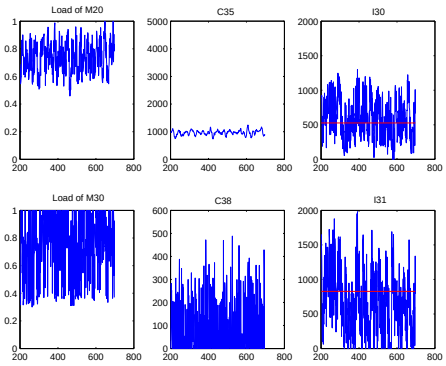
In this case, the split average is larger than the average of $\frac{D}{E}$. So there are more products to meet the demand for D than to meet the demand for E at steady state. In a deterministic setting, all of the demands and targets can be met. Since the split in $M20$ does not make enough products to meet the demand for E, some of the fast devices in $I20$ are used to make the slow devices through $M20$ to meet the demand for E.

Stochasticity is introduced in the throughput times, yields and customer demands. At first, no move suppression is applied on manipulated variables; all controlled variables are equally weighted at 1. As shown in Figure 5, due to the aggressive responses, the components warehouse inventories for both fast and slow devices are depleted so often that many backorders are generated. The WIPs for $M20$ and $M30$ exceed the capacity in most of the time during the simulation, and not enough products needed can be produced in this situation.

In order to smooth the responses and achieve robustness, move suppressions of [10 10 10 0 10] are applied on manipulated variables [C35 C36 C37 C38 C39]. Only $C38$ has zero move suppression because we want this variable to respond to the demand as fast as possible. The inventory levels are still equally weighted at 1. The results in Figure 6 show the starts are smoother than



(a) With no move suppression



(b) With move suppression

Figure 5: Selected responses in Assembly/Test2 stochastic split problem Case 2

before. The variance on $C35$ decreases 99% compared to the previous case without move suppression. The very noisy response of the WIP signals is mainly due to the high stochasticity in manufacturing nodes shown in

Figure 3 which cannot be changed by tuning. However, compared with the case with no move suppression, the variance on $M20$ WIP decreases 90.5% and for $M30$ decreases 73.6%. The inventory levels are smoother and high enough to meet the customer demands in most cases. Furthermore, the variance on $I30$ is reduced by 43%. We only observe a little bit of backorders for the demand for E although the split is not balanced, because we can take advantage of $C38$ to transfer some fast devices from $I20$ to $I31$. Although the WIP still reaches the capacity limit on occasion, much less backorders are generated compared with the case using no move suppression. Figure 6 shows the comparison of the backorders generated in each case. The unfilled fast device order shown on right side in Figure 6 decreases 90.5% by applying move suppression on manipulated variables and the unfilled slow device order decreases 73.6%. For a more detailed comparison, refer to [5]. Clearly, the performance and robustness are improved by applying move suppression to the problem.

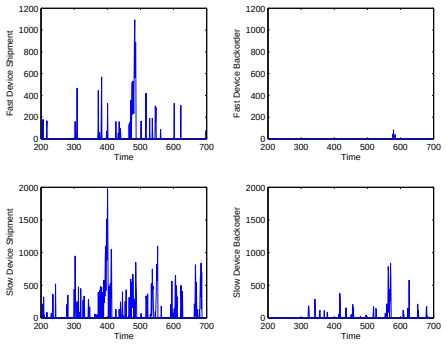


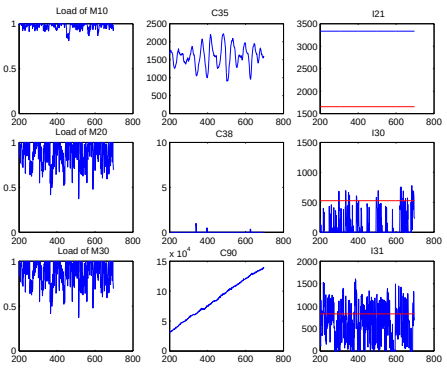
Figure 6: Backorders in Assembly/Test2 stochastic split problem Case 2: left side: with no move suppresion; right side: with move suppression

3.2 Case 3 (Discard Low)

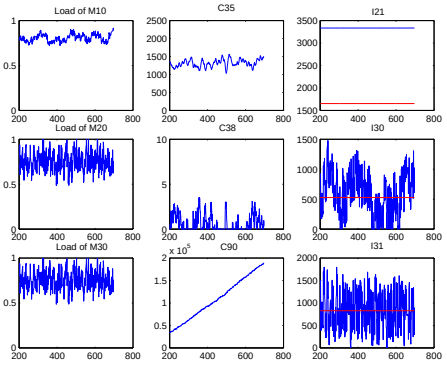
In Case 3, the average split is smaller than the average of $\frac{D}{E}$. At steady state, there will be not enough items to meet the demand for D if demand for E is met without any excessive products left in $I21$ or $I31$. Even in a deterministic setting (not shown), in order to meet the demand for D, the total amount of items processed in $M10$ and $M20$ must be increased. While enough items will be shipped to $I20$ and $I30$ to meet the demand for D, but there will be more than enough products to satisfy the demand for E. The inventory will be held until it reaches the high limit of capacity and excessive products will be scrapped due to the limited capacity of $I21$. Although no backorder is generated, the loads of F/T1 and A/T2 exceed 95%.

When stochasticity is introduced in the throughput times, yields of the manufacturing nodes and two cutomer demands, we first try to test the original capacity which is 45000, 7500, 2500 for $M10$, $M20$ and $M30$ respectively. In order to achieve a smooth response, move suppressions of [100 100 1000 0.1 100] are used. The total amount of products starting from $C35$ have to increase to produce enough high speed devices and excess low speed devices. Consequently, the inventory $I21$ needs to store the excess slow devices and scrap when it reaches the limit. Therefore, the output weight for this variable is set to be zero. The results are shown in Figure 7. The inventory level of $I21$ stays at the capacity limit all the time and many slow devices are scrapped though $C90$. The loads of $M10$ and $M20$ exceed the capacity most of the time and not enough products can be produced on time with these capacity settings. Many backorders are observed becuase the inventories of $I30$ and $I31$ are depleted. This is a result of the limited capacity and can not be overcome simply by controller tuning. The only way to solve this problem is to allocate capacity for each manufacturing node which is sufficient to meet the demands.

If we increase the capacities of $M10$ and $M30$ by 20% and $M20$ by 30% and apply the same move suppressions, the performance is significantly improved as shown Figure 7. One can find here the loads of $M10$, $M20$ and $M30$ are all below 100%. We have enough capacity to process items as many as they are needed. The inventory levels can track the targets without being depleted so often. The variance of the start is smaller than that in previous case with limited capacity. For instance, the variance of $C35$ is reduced by 87%. $I21$ still reaches the high inventory limit and many slow devices are scrapped through $C90$ in order to increase the amount of items for meeting the demand for D. The inventories of $I30$ and $I31$ are high enough to meet the demands, although sometimes they are still depleted due to the stochasticity. As shown in right side of Figure 8, the unfilled order for slow devices is almost zero with enough capacity. The unfilled order for fast devices is reduced by 93.7% compared to the limited capacity case as shown in left side of Figure 8. Although the slow devices are more than enough, some fast devices are still reconfigured to meet the slow devices demand due to the uncertainty in the manufacturing processes and the customer demands. Since we lose the flexibility of using $C38$ to make some fast devices from $I21$ and the split can not make enough fast devices, the backorders of the demand for D are more than that for E. This simulation study shows it necessary to have an external decision policy that allocates capacity based on longer-term demand information and economic considerations.



(a) Limited capacity



(b) Increased capacity

Figure 7: Seleced responses inAssembly/Test2 stochastic split problem Case 3

4 Conclusions

A Model Predictive Control formulation as a tactical controller in semiconductor manufacturing is presented and validated via simulation. Although relying on a linear model with fixed throughput times and yields, MPC can have satisfactory performance for systems with high stochasticity and uncertainty. The flexibility to handle constraints makes it possible to track inventory targets while meeting customer demand. Move suppression in MPC plays an important role in achieving robustness under uncertainty in systems. Increasing move suppression can help make the responses smooth with less backorders. It can also influence WIP in factories to not change sharply, which is desirable in practice. Both the inventory targets and capacity limits should be provided by an external algorithm. Enough inventories should be held to buffer the stochasicity on both demand side and supply side. Sufficient capacity makes enough products be produced on time to meet the customer demand. The actions of the MPC controller can show when and where the capacities are

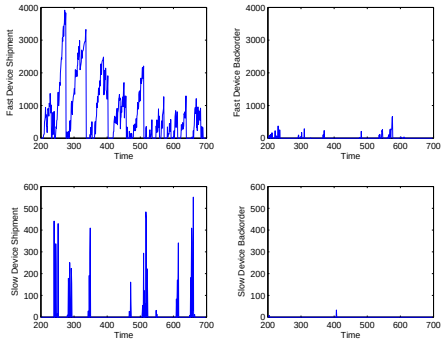


Figure 8: Backorders in Assembly/Test2 stochastic split problem Case 3: left side: limited capacity; right side: increased capacity

depleted. These also give a reasonable justification for expanding capacity and insights into the interaction between the MPC decision policy and additional external decision policies.

5 Acknowledgement

Support from the Intel Research Council for this research is gratefully acknowledged.

References

[1] Gunasekaran, A and E.W.T. Ngai. Information systems in supply chain integration and management *European Journal of Operational Research*, to appear, 2003.

[2] K. G. Kempf. Control-Oriented Approaches to Supply Chain Management in Semiconductor Manufacturing *American Control Conference,* Boston, MA, 2004.

[3] M. W. Braun, D. E. Rivera, M. E. Flores, W. M. Carlyle and K. G. Kempf. A Model Predictive Control framework for robust management of multi-product, multi-Echelon demand networks, *Annual Reviews in Control, Special Issue on Enterprise Integration and Networking*, Vol.27, Issue 2, pp. 229-245, 2003.

[4] W. Wang, D. E. Rivera and K. G. Kempf. Centralized Model Predictive Control Strategies for Inventory Management in Semiconductor manufacturing Supply Chains, *Proc. American Control Conference*, Denver, CO, June 2003, pp. 585-590.

[5] W. Wang, J. Ryu, D. E. Rivera, K. G. Kempf and K. D. Smith. A Model Predictive Control Approach for Managing Semiconductor Manufacturing Supply Chains under Uncertainty, *Annual AIChE Meeting* , paper 446d: pgs 1-34, San Francisco, CA, Nov 16-21, 2003.