

A Novel Model Predictive Control Algorithm for Supply Chain Management in Semiconductor Manufacturing

Daniel E. Rivera

Arizona State University

Hans D. Mittelmann

Arizona State University

Hessam S. Sarjoughian

Arizona State University

Karl G. Kempf

Intel Corporation

Abstract: Supply chains in semiconductor manufacturing are characterized by integrating dynamics, nonlinearity and high stochasticity. In this paper, we present a novel Model Predictive Control (MPC) algorithm for Supply Chain Management (SCM) in semiconductor manufacturing. A Type II filter is designed to attenuate the integrating noise such as that exhibited by unforecasted customer demand. The selection of the filter gain provides the flexibility to achieve better performance and robustness. The forecast of customer demand plays a critical role in the algorithm. The advantages of this novel MPC algorithm are demonstrated through case studies of a representative supply chain problem in semiconductor manufacturing which involve scenarios of customer demand forecast error and anticipated periodic demand.

1. Introduction: Supply chain management (SCM) has gained significance as one of the 21st century manufacturing paradigms for improving organizational competitiveness. SCM has been considered as a competitive strategy for integrating supplies and customers with the objective of improving responsiveness and flexibility of manufacturing organizations [1]. SCM is also vital for discrete parts manufacturing such as semiconductor manufacturing. It integrates factories, warehouses and

customers for physical flow of materials, manufacturing planning and control, and physical distribution. The role of integrated decision policies for improving supply chain management in the semiconductor industries is described in some detail in (Kempf, 2004)[4].

Model Predictive Control (MPC) is a control strategy that has been widely used in chemical process industries. Recently, the work of Braun *et al.* [5] applies a partially decentralized MPC structure for inventory control in supply chain. A centralized MPC strategy is successfully implemented for semiconductor manufacturing SCM to track inventory targets and satisfy stochastic customer demands given uncertainty on throughput time and yield in factories [6, 7]. Good system performance and robustness are achieved using judicious choices of move suppression. In this paper, a state estimation based MPC is formulated to handle the short time scale stochasticity in semiconductor manufacturing and make optimal daily decisions for supply chain management. A Type II filter addresses the output responses individually, as opposed to relying on move suppression to slow down the manipulated variables. The anticipation of the measured disturbance, i.e. forecasted customer demand, improves system performance by making the system respond

to demand changes in a more efficient way.

The paper is organized as follows. In Section 2, the state estimation-based Model Predictive Control algorithm is discussed. In Section 3, a representative problem in semiconductor manufacturing is studied. The effects of filter gain tuning and anticipation of customer demand are demonstrated with examples involving customer demand forecast error and periodic demand. The paper concludes with a discussion of the flexibility and advantages of using MPC in SCM.

2. Model Predictive Control: The MPC controller algorithm implemented in this work is explained in this section.

2.1 Controller Model: The MPC controllers considered in this paper rely on a linear state-space model [2]:

$$\begin{aligned} x(k) &= Ax(k-1) + B_u u(k-1) + B_d d(k-1) \\ y(k) &= Cx(k) + D_d d(k) \end{aligned} \quad (1)$$

The measurement vector y_m is described as:

$$y_m(k) = y(k) + v(k) \quad (2)$$

The inputs are u , d and v representing manipulated inputs, load disturbance and measurement noise respectively. y is the noise free output and y_m is the measured inventory levels or controlled variable. The other corresponding variables in supply chain system are the starts of factories (u) and customer demand (d) which is treated as a measured disturbance with anticipation. We assume that d is a stochastic signal described through the following model:

$$\begin{aligned} x_w(k) &= A_w x_w(k-1) + B_w w(k-1) \\ d(k) &= C_w x_w(k) \end{aligned} \quad (3)$$

where A_w has all the eigenvalues inside the unit disk and $w(k)$ is a vector of integrated white noise. Augmenting (1) with (3) and taking the difference form gives:

$$\begin{aligned} X(k) &= \Phi X(k-1) + \Gamma_u \Delta u(k-1) + \Gamma_w \Delta w(k-1) \\ \hat{y}(k) &= \Xi X(k) + v(k) \end{aligned} \quad (4)$$

where

$$X(k) = \begin{bmatrix} \Delta x(k) \\ \Delta w(k) \\ y(k) \end{bmatrix} \quad (5)$$

$$\Phi = \begin{bmatrix} A & B_d C_w & 0 \\ 0 & A_w & 0 \\ CA & CB_d C_w + D_d C_w A_w & I \end{bmatrix} \quad (6)$$

$$\Gamma_u = \begin{bmatrix} B_u \\ 0 \\ CB_u \end{bmatrix} \quad (7)$$

$$\Gamma_w = \begin{bmatrix} 0 \\ B_w \\ D_d C_w B_w \end{bmatrix} \quad (8)$$

$$\Xi = [0 \ 0 \ I] \quad (9)$$

Here $\Delta * (k) = *(k) - *(k-1)$.

2.2 State Estimation and Prediction: The states of the process model and the unmeasured disturbance and noise must be estimated using the augmented system model in (4). A Kalman filter can be used for the optimal state estimator as follows [3].

$$X(k|k-1) = \Phi X(k-1|k-1) + \Gamma_u \Delta u(k-1) \quad (10)$$

$$X(k|k) = X(k|k-1) + K_f (y_m(k) - \Xi X(k|k-1)) \quad (11)$$

Here K_f is the filter gain usually found by solving the algebraic Riccati equation. However, the covariance matrices of load disturbance and measurement noise may not be accurately known. It is more practical to set the filter gain as a tuning parameter based on the signal-to-noise ratio. Since we do not need the prediction of each state in system, it is convenient to lump the effect of all disturbances on the outputs only (i.e. $B_d = 0$, $D_d = I$). If we assume the load disturbance is double integrated white noise (because of the integrating dynamics in the supply chain) and no information is available on the correlation of disturbances among different outputs (i.e. $A_w = I$, $B_w = I$, $C_w = I$) [2], then (4) becomes:

$$\begin{aligned} \begin{bmatrix} \Delta x(k) \\ \Delta x_w(k) \\ y(k) \end{bmatrix} &= \begin{bmatrix} A & 0 & 0 \\ 0 & I & 0 \\ CA & I & I \end{bmatrix} \times \begin{bmatrix} \Delta x(k-1) \\ \Delta x_w(k-1) \\ y(k-1) \end{bmatrix} \\ &+ \begin{bmatrix} B_u \\ 0 \\ CB_u \end{bmatrix} \Delta u(k-1) + \begin{bmatrix} 0 \\ I \\ I \end{bmatrix} \Delta w(k-1) \end{aligned} \quad (12)$$

where Δw is white noise with following covariance matrix:

$$E\{\Delta w \Delta w^T\} = \text{diag}\{q_1, \dots, q_{n_y}\} \quad (13)$$

We also assume that the measurement noise is white with following covariance matrix:

$$E\{vv^T\} = \text{diag}\{r_1, \dots, r_{n_y}\} \quad (14)$$

where n_y is the number of outputs. For open-loop stable systems, it can be shown that the Type II optimal filter gain K_f for the system is also parameterized in terms of an n_y -dimension real vector with each element lying in $(0,1]$ [2].

$$K_f = \begin{bmatrix} 0 \\ F_b \\ F_a \end{bmatrix} \quad (15)$$

where

$$\begin{aligned} F_b &= \text{diag}\{(f_b)_1, \dots, (f_b)_{n_y}\} \\ F_a &= \text{diag}\{(f_a)_1, \dots, (f_a)_{n_y}\} \\ (f_b)_i &= \frac{(f_a)_i^2}{2 - (f_a)_i} \text{ for } 1 \leq i \leq n_y \end{aligned} \quad (16)$$

and

$$\begin{aligned} (f_a)_i &\rightarrow 0 \text{ as } q_i/r_i \rightarrow 0, \\ (f_a)_i &\rightarrow 1 \text{ as } q_i/r_i \rightarrow \infty \end{aligned} \quad (17)$$

The first term in (11), $X(k|k-1)$, includes all of the deterministic information such as the nominal system delay and measured disturbance anticipation. The second term is the prediction error generated by the stochasticity and uncertainty in the system. As f_a approaches zero, the system ignores most of the prediction error and the solution is mainly determined by the deterministic model and the anticipation. The system will compensate all of the prediction error from the stochasticity and uncertainty if f_a is one.

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

$$\min_{\Delta u(k|k) \dots \Delta u(k+m-1|k)} J \quad (18)$$

where the individual terms of J correspond to:

$$\begin{aligned} &\underbrace{\sum_{\ell=1}^p Q_e(\ell) (\hat{y}(k+\ell|k) - r(k+\ell))^2}_{\text{Keep Inventories at Inventory Planning Setpoints}} \\ J &= \sum_{\ell=1}^p Q_e(\ell) (\hat{y}(k+\ell|k) - r(k+\ell))^2 \\ &\quad \underbrace{\sum_{\ell=1}^m Q_{\Delta u}(\ell) (\Delta u(k+\ell-1|k))^2}_{\text{Penalize Changes in Starts}} \\ &+ \sum_{\ell=1}^m Q_{\Delta u}(\ell) (\Delta u(k+\ell-1|k))^2 \quad (19) \\ &\quad \underbrace{\sum_{\ell=1}^m Q_u(\ell) (u(k+\ell-1|k) - u_{target}(k+\ell-1|k))^2}_{\text{Maintain Starts at Strategic Planning Targets}} \end{aligned}$$

subject to the capacity constraints on the starts, the change rate of starts, the warehouse inventories and factory Work-In-Progress. 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.

3. Case Study: In this Section, we present a basic supply chain problem which includes distinguishing features of semiconductor manufacturing. The basic semiconductor manufacturing process is shown in Figure 1. Wafers are first fabricated and tested in

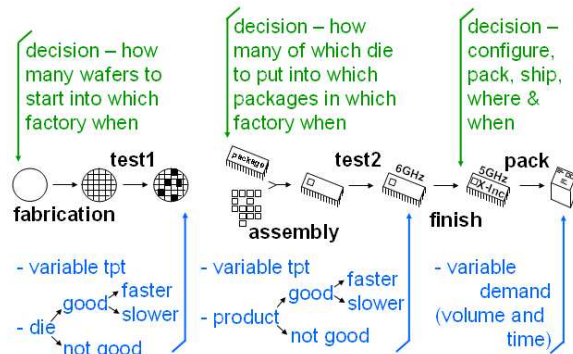


Figure 1: Sequence of steps in semiconductor manufacturing

Fab/Test1. Transistors are built on a silicon wafer and then interconnected to form circuits in a fabrication process. In Assembly/Test2 process, the individual die are cut from the wafers and mounted in

packages to protect them. The semi-finished goods are configured, packed and shipped to customers in the finishing manufacturing stage. One wafer can be processed and configured to make different products. A fluid representation of a three-node semiconductor manufacturing supply chain (consisting of one Fab/Test1, one Assembly/Test2, and one finish node) and its corresponding inventory locations is shown in Figure 2. In a very general sense, the manufacturing nodes are represented as “pipes”, while the inventory locations are represented as “tanks”; material in these pipes and tanks correspond to factory Work-in-Progress (WIP) and warehouse inventory, respectively. A discrete time model is used to describe the

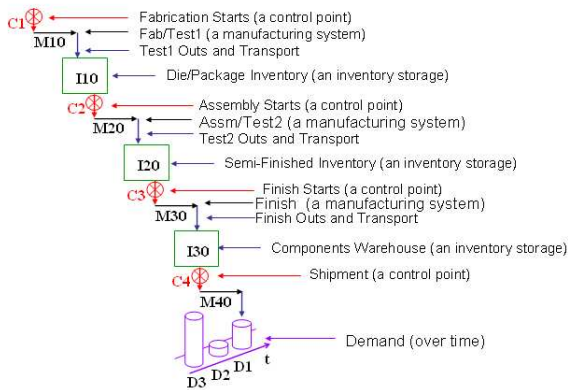


Figure 2: Fluid representation of a three node semiconductor mfg. supply chain

dynamics of supply chain. For each inventory position at day k , we can develop the following equation based on material balance.

$$\begin{aligned}
 I_i(k+1) &= I_i(k) + Y_j C_j(k - \theta_j) - C_{j+1}(k) \\
 i &\in \{10, 20, 30\} \\
 j &\in \{1, 2, 3\}
 \end{aligned} \tag{20}$$

where I_i is the inventory stored in position i , C_j is the start of factory j and C_4 is the anticipation of future customer demand; Y_j and θ_j are the yield and throughput time of factory j respectively. These models could be organized in state-space form and used as nominal controller model. Both the throughput time and yield are deterministic in the nominal controller model, while the stochasticity is introduced on these parameters in the simulation model as described in Table 1. Many requirements of SCM in

Factory	nodes	M10	M20	M30	M40
load \in [0%,70%]	Min TPT (days)	30.0	5.0	1.0	1.0
	Ave TPT (days)	32.0	6.0	2.0	1.0
	Max TPT (days)	34.0	7.0	3.0	1.0
load \in (70%,90%]	Distribution	Unif	Unif	Unif	
	Min TPT (days)	32.0	5.0	1.0	1.0
	Ave TPT (days)	35.0	6.0	2.0	1.0
load \in (90%,100%]	Max TPT (days)	38.0	7.0	3.0	1.0
	Min TPT (days)	35.0	5.0	1.0	1.0
	Ave TPT (days)	40.0	6.0	2.0	1.0
Yield	Max TPT (days)	45.0	7.0	3.0	1.0
	Min %	93.0	98.0	98.5	100.0
	Ave %	95.0	98.5	99.0	100.0
Distribution	Max %	97.0	99.0	99.5	100.0
	Distribution	Unif	Unif	Unif	
	Capacity	Max Items	4.5E4	7500	2500
Inventory	nodes	I10	I20	I30	
	UCL(Items)	1.2E4	6E3	3E3	
	TAR(Items)	5706	2853	1427	
	LCL(Items)	1000	1000	1000	
	Max(Items)	2E4	1E4	1E4	

Table 1: Manufacturing and inventory nodes data for basic problem with backlog: TPT refers to throughput time; Unif to uniform distribution; UCL to Upper Control Limits, TAR to Target, LCL to Lower Control Limits

semiconductor manufacturing are more properly expressed as constraints on the process variables. There are three types of constraints that can be imposed in SCM as follows:

- *Manipulated variable constraints:* these are hard high and low limits on the starts of factories due to the capacity

$$\begin{aligned}
 C_j^{min} &\leq C_j(k) \leq C_j^{max} \\
 j &\in \{1, 2, 3\}
 \end{aligned} \tag{21}$$

where C_j^{min} , C_j^{max} are the high and low limits of the start to factory j . These limits can be constants or time dependent.

- *Manipulated variable rate constraints:* these are hard limits on the maximum and minimum move size of the acceptable thrash of factory starts

$$\begin{aligned}
 \Delta C_j^{min} &\leq \Delta C_j(k) \leq \Delta C_j^{max} \\
 j &\in \{1, 2, 3\}
 \end{aligned} \tag{22}$$

where ΔC_j^{min} , ΔC_j^{max} are the high and low limits of the start variation to factory j . They can also be constants or vary over a horizon.

- *Output variable constraints:* inventory levels have specified targets and the objective function

minimizes their deviations from the setpoints. Besides the setpoints, they also have high and low limits on the inventories. To avoid infeasible solutions, the constraints on controlled variables are treated as soft constraints in the objective function.

$$I_i^{min} \leq I_i(k) \leq I_i^{max} \quad (23)$$

$$i \in \{10, 20, 30\}$$

The customer demand is treated as a measured disturbance with anticipation. It is kept for one week, and then it is lost. The error between the actual demand and the forecast could be significantly large. Based on different customer demands and anticipation, we develop two scenarios to study the effects of the filter gain and anticipation.

3.1 Scenario 1: Stationary Demand and Forecast Error In this case, we demonstrate the effects of Type II filter gain to handle forecast error in customer demand. The actual customer demand and two different demand forecasts are shown in Figure 3. The first forecast is ten day moving average of the actual demand, while the variance of the actual demand is larger than that of the anticipation. If this forecast is used, the results as shown in Figure 4 demonstrate that the starts are similar to the demand

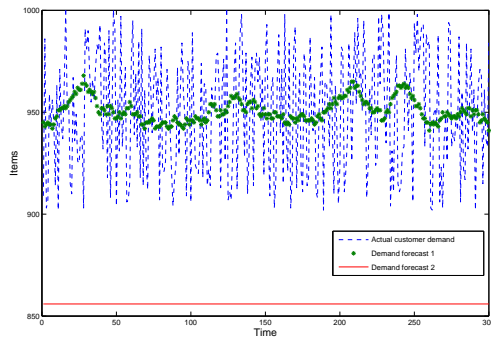


Figure 3: Stationary customer demand and forecast error in scenario 1: dashed: actual customer demand, star: forecast with the same average of actual demand, dot: forecast with 100 units error to the average of actual demand.

forecast and the oscillation on the inventory is mainly caused by the stochastic throughput time and yield in

the semiconductor manufacturing process. No backlog is generated and the simulation also shows the inventories are more than enough to meet the customer demand. Compared with the actual customer demand, the starts response for each of the factory is quite smooth.

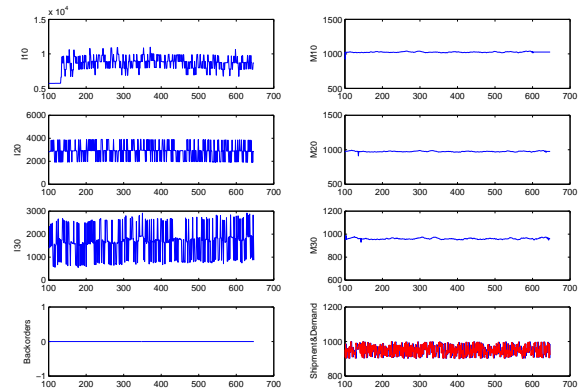


Figure 4: System responses using the same average forecast with $f_a = 0$, Scenario 1

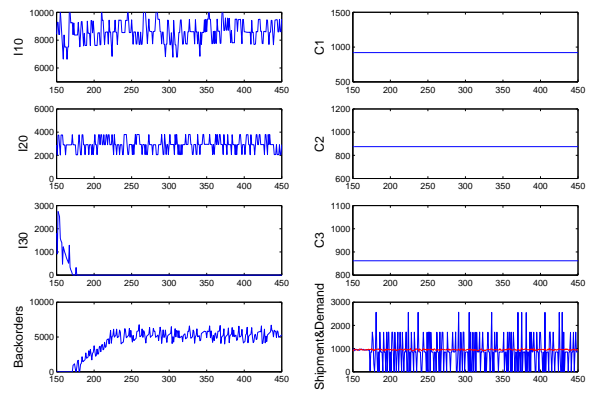


Figure 5: System responses using an erroneous forecast with $f_a = 0$, Scenario 1

With an erroneous demand forecast as shown in Figure 3, the average of the anticipation signal is 100 units smaller than that of the actual customer demand. If the move suppression is zero, the output weight is 1 for each controlled variable and the filter gain is zero on each output, the controller ignores the

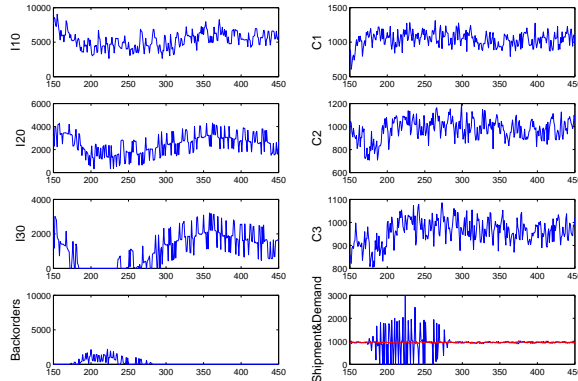


Figure 6: System responses using erroneous forecast with $f_a = 0.04$, Scenario 1

average difference between the actual demand and the forecast and tries to follow the anticipation only. Warehouse inventory $I30$ is depleted and many backorders are accumulated since the controller will not compensate the prediction error, however the starts ($C_i, i = 1, 2, 3$) are very smooth as shown in Figure 5.

If the filter gain f_a is increased to 0.04, the backorders are dramatically reduced while the variance on the starts increases, as shown in Figure 6. The starts begin to increase after the controller measures the actual demand and find the difference between the measurement and the anticipation. The rising speed is determined by the filter gain. The larger the filter gain, the faster the controller increases the starts. If the speed of response is not sufficiently high, the safety stock in inventory $I30$ is depleted and backorders are generated.

As shown in this case, increasing filter gain can make the controller compensate the forecast error. The larger the filter gain, the faster the controller can compensate the error. However, if the filter gain is too large, the responses will be very noisy since the controller tries to chase the noise from the customer demand and stochastic manufacturing process and this also generates backorders. The relationship between the backorders, the variance of the Fab starts and the filter gain for this case can be shown in Figure 7. There is always a tradeoff between the response speed and robustness, either of which will influence the total cost of inventory holding, revenue and backlog. The optimal filter gain in this scenario $f_a \cong 0.05$

as described in Figure 7.

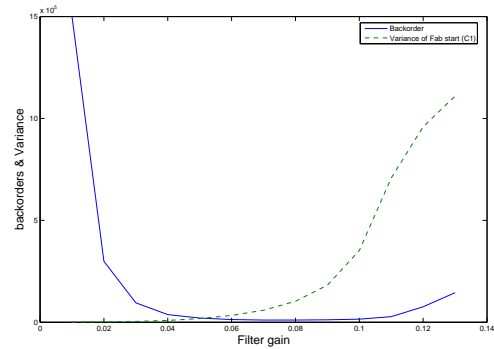


Figure 7: Tradeoff between the filter gain and cumulative backorders (solid line), and Fab/Test starts variance (dashed line)

3.2 Scenario 2: Periodic Customer Demand

In practice, customer demand may change dramatically from day to day, while operations may want the starts of the factories to stay as constant as possible. In this case, we apply a sinusoidal customer demand to test the effects of anticipation in the demand forecast and to evaluate the flexibility of this MPC algorithm. The customer demand is a stochastic sinusoidal signal with the mean of 950 units, amplitude of 100 units and frequency of 0.1 rad/sec. The first demand forecast is a deterministic sinusoidal signal with the same mean, amplitude and frequency as the stochastic demand and the second demand forecast is the average of the stochastic demand. In both cases, the move suppression and output weight parameters are zero and one respectively. The filter gain is set to be 0.01 to achieve the robustness. As shown in Figure 8, for this first demand forecast, there is no backorder, and the inventory has low variance and no offset with the target. However, the starts change sinusoidally just like the actual demand; this may not be desirable to operations. If the stationary demand forecast is applied, the results are shown in Figure 9. No backorders are generated and the starts are smooth like the forecast, while the closest inventory $I30$ varies periodically corresponding to the customer demand and absorbs the variance of the demand.

4. Conclusions: A novel Model Predictive Control formulation is presented for supply chain man-

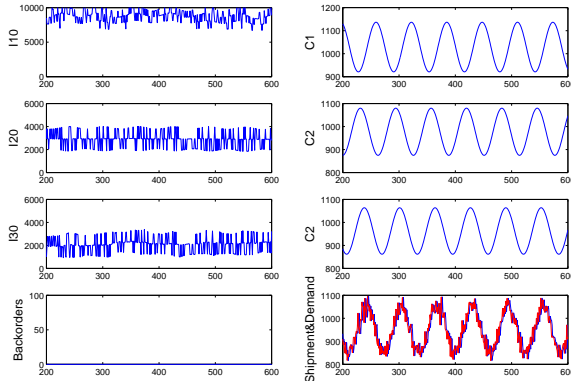


Figure 8: System responses with periodic demand forecast, Scenario 2: $f_a = 0.01$

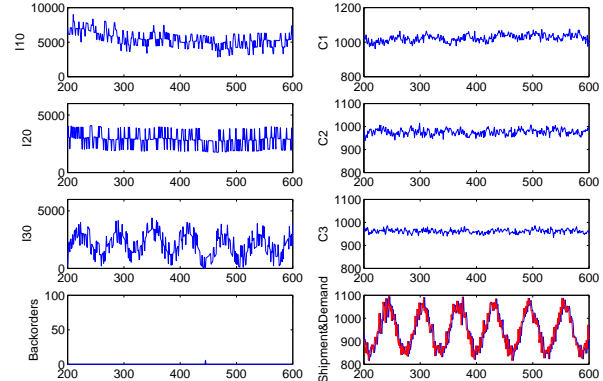


Figure 9: System responses with stationary demand forecast, Scenario 2: $f_a = 0.01$

agement in semiconductor manufacturing. The selection of proper tuning parameter helps the system to achieve both robustness and performance with the linear deterministic nominal model despite the stochasticity and nonlinearity in the plant. In most cases, the choice of a small filter gain improves robustness. However, if there is a large error between the average of customer demand and the forecast, a larger filter gain has to be used to make the controller compensate for the error sufficiently fast. There is always a tradeoff to pick the proper filter gain. The anticipation of the measured disturbance is critical for achieving better performance. Proper anticipation can help the system to meet different requirements with no backlog cost. Simulation results demonstrate the effects of both filter gain and anticipation. From these we see that MPC as a flexible powerful tool for making daily decision for SCM in semiconductor manufacturing, in the presence of high stochasticity and nonlinearity.

5. Acknowledgement: Support for this research from the Intel Research Council and the National Science Foundation (grant DMI-0432439) is gratefully acknowledged.

References

[1] Editorial. Supply chain management: Theory and application *European Journal of Operational Research*, 159 (2004) 265-268.

[2] Lee, J and Z. H. Yu. Tuning of model predictive controllers for robust performance *Computers chem. Engng*, Vol. 18, No. 1, pp. 15-37, 1994.

[3] Åström, K. J. and B. Wittenmark, *Computer Controlled Systems: Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[4] K. G. Kempf. Control-Oriented Approaches to Supply Chain Management in Semiconductor Manufacturing, *Proc. American Control Conference*, Boston, MA, June 20-July 2 2004, pp. 4563-4576.

[5] 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.

[6] 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.

[7] W. Wang, D. E. Rivera, K. G. Kempf and K. D. Smith. A Model Predictive Control Strategy for Supply Chain Management in Semiconductor Manufacturing under Uncertainty, *Proc. American Control Conference*, Boston, MA, June 20-July 2 2004, pp. 4577-4582.