Short and Long-term Scheduling in Semiconductor Manufacturing

M.C. Colantonio*, L. Papageorgiou# , N. Shah
Centre for Process Systems Engng., Imperial College
London SW7 2BY, UK
#Dept. of Chem. Engng, University College London
Torrington Place, London WC1E 7JE, UK

Abstract

This paper addresses the scheduling problem in semiconductor manufacturing. A two level hierarchical structure is considered to take into account different horizons in the decision making process. Long-term planning is solved by means of an $l_1 - norm$ Model Predictive Controller which gives the release policy to a short-term scheduler. The latter is based on a State-Task-Network representation of the batch recipe and provides the detailed operation of the fab.

1 Introduction

A semiconductor manufacturing process is typically large-scale, featuring different steps where many layers are made separately by printing a chemical pattern over silicon wafers. An important structural characteristic of this discrete-event process, imposed by huge machine costs, is the reentrant line. This means that the wafers may return to the same machine for processing at different steps and may also spend some time waiting to be processed. It results in large cycle times which lead to low production volume. The operation of the fab involves multiple product accommodation by a single production line, restrictions on machine availability, regular maintenance and wafer wait-time constraints.

Although the manufacturing floor has been automated, it is necessary to provide new operational procedures in order to reduce unpredictable behaviour and to improve performance. Then, the aim of a process controller is to reduce cycle time while keeping inventories low and maximising production. The size and complexity of the system makes it natural to divide the planning/scheduling problem into levels of hierarchy characterized by the horizon of planning and the data required in the decision process (Ed Adl *et al.*, 1996). We will consider a hierarchical structure with two level decision making: a long-term horizon for campaign planning and a short-term horizon scheduling at the level of the process operation (Fig. 1). Long-term planning is based on reliable demand predictions over a long period of time and has the advantage of minimising the number and cost of changeovers. Short-term production planning involves determining how the resources of the plant will best be utilised over a period of time to meet a specified objective, typically demand and stock build requirements.

Vargas-Villamil *et al.* (1997) have proposed the extension of Model Predictive Control (MPC) concepts to address long-term scheduling in semiconductor manufacturing lines. An l_1-norm finite moving horizon cost function is considered in a state-space formulation of the MPC via

---

*Author to whom all correspondence should be addressed. e-mail: c.colantonio@ic.ac.uk

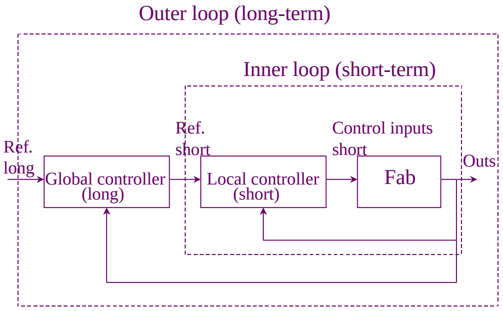

Figure 1: Hierarchical Structure

an aggregated linear model. This formulation for the MPC is attractive in the sense that the optimisation problem can be transformed into a linear programming (LP) one. An open-loop MPC controller/scheduler provides the starts and utilisation targets for a short-term controller that uses a pull-like policy, tracking utilisation targets (Vargas-Villamil *et al.*, 1998). A small time scale modeling concept to approximate an asynchronous system to a synchronous one (Tsakalis *et al.*, 1997) is used in order to avoid dealing with the complexity of discrete-event system model itself (El Adl et al, 1996). But, as schedule optimisation has become an important factor for improving performance through better utilisation of the available resources, researchers are focused on solving all aspects related to deal with the aforementioned complexity. Improved understanding of the formulation and solution of Mixed-Integer-Programming as well as advances in computer hardware, allowed more general formulations to be developed and larger problems to be tackled. Pierce and Realff (1996) addressed the solution of the scheduling problem for multi-chip fabrication adopting a State-Task-Network (STN) (Kondili *et al.*, 1993) to represent the batch recipe and an optimisation model to produce a Mixed Integer Linear Programming (MILP) formulation, assuming fixed process conditions.

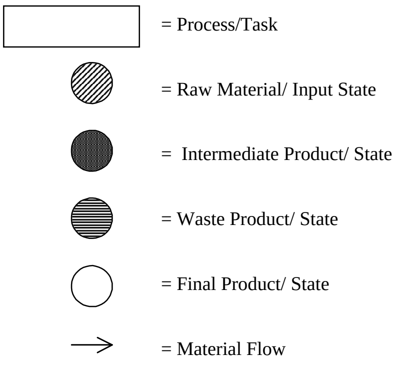

Figure 2: STN representation

The key advantage of a STN (Fig. 2) over a 'task-only' network is its ability to represent unambiguously processes with shared intermediates, recycles of materials and those with more than one different processing route leading to the same intermediate or final products. However, as the solution of the MILP can be computationally intensive, the STN is ideal for problems with relatively short time horizons as those encountered in short-term scheduling.

During the last decade, an advanced integrated software tool known as gBSS (Papageorgiou *et al.*, 1992) has been developed for optimal planning and scheduling of batch plants. It is based on a STN description of the process and a general characterisation of resources. It also adopts a representation of the problem related to the discrete-event system that shields the users from the resulting complex mathematical models and solution procedures. These models are automatically generated and used to solve the underlying MILP problem. As a result, a detailed production plan satisfying constraints and orders is obtained.

In this paper, we consider the implementation of the two level hierarchical structure as follows: a MPC formulation similar to the one proposed in Vargas-Villamil *et al.* (1997) provides the start release policy to a short-term scheduler in gBSS. An objective function which results in a numerically more robust optimisation problem for the MPC is proposed. We pose the solution in closed-loop form to address the performance of the complete hierarchical structure in the presence of plant/model mismatch.

In Section 2, we describe the main aspects of the proposed MPC and its formulation as the solution of an LP problem. Section 3 contains a brief description of the short-term scheduler. Results of the application of the proposed methodology to the five-machine six-steps semiconductor process described in Vargas-Villamil (1998) are presented is Section 4. Main conclusions are outlined in Section 5.

2 Long-Term Planning

The global dynamics of fabs involving different time scales can be approximated in the long-term using an aggregated model, where the interconection matrices are average values of different effects. These include cumulative effect of transportation, batching, lags, utilisation and buffer constraints, maintenance, etc. Then, the model of the fab can be represented in the long-term by the following discrete-time linear system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + B_d d_k && \text{(1a)} \\ y_k &= Cx_k \; , \; x(0) = x_0 && \text{(1b)} \end{aligned}$$

subject to:

$$\begin{aligned} hy_k^{min} &\leq H\left(x_k, y_k, u_k\right) \leq \hat{h}y_k^{max} && \text{(2a)} \\ gu_k^{min} &\leq G\left(x_k, y_k, u_k\right) \leq \hat{g}u_k^{max} && \text{(2b)} \end{aligned}$$

where $x_k \in R^n$, $u_k \in R^m$, $d_k \in R^d$ and $y_k \in R^l$. $x_{i,k}[lots]$ are the inventories in the i-th buffer, waiting to be processed at time k $(i = 1, ..., n-1)$, and $x_{n,k}[lots/shift]$ is the throughput volume. $u_{j,k}$ are the utilisation targets of the j-th step $(j = 1, ..., m-1)$ and $u_{m,k}[lots]$ are the starts entering the fab. The wafers produced are given by $y_{i,k}$ for $i = 1, ..., l-1$ while $y_{l,k}$ is the throughput volume. H and G are constraint transition matrices on the outputs and inputs, respectively.

For the 5-machines-6-steps process considered in Vargas-Villamil *et al.* (1998), matrices in Eqs. 1a and 1b are defined by:

$$\begin{aligned} A &= \begin{bmatrix} \mathbf{I}_{6\times6} & 0 \\ 0 & 0 \end{bmatrix}; B = \begin{bmatrix} -R_1 & 0 & 0 & 0 & 0 & 0 & 1 \\ R_1 & -R_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & R_2 & -R_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & R_3 & -R_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_4 & -R_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & R_5 & -R_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & R_6 & 0 \end{bmatrix} \\ B_d &= [0] \quad ; \quad C = [\mathbf{I}_{7\times7}] \end{aligned} \tag{3}$$

being $R_i[lots/shift]$ the effective machine production rates for the i-th step.

In this case, constraints in Eqs. 2a and 2b are specified as follows:

$$u_k \geq 0 \tag{4}$$

$$y_k^{\min} \leq y_k \leq y_k^{\max} \tag{5}$$

$$C_{Av}^{\min} \leq C_L u_k \leq C_{Av}^{\max} \tag{6}$$

where Eq. 4 are control input sign constraints, Eq. 5 are maximum and minimum capacity of inventories and Eq. 6 are availability constraints. C_L is a permutation matrix related to the reentrant nature of the process.

2.1 MPC formulation

MPC is an optimal control based method to select control inputs by minimising an objective function defined in terms of both present and future process outputs (Garcia et al, 1995; Henson and Seborg, 1997). The control problem is solved as follows: with knowledge of the current output y_k, we seek for a control that minimises the objective and implement only the first control move. When a new measure becomes available, the parameters of the problem are updated and the optimisation is carried out again. The solution provides the next control move. This repeated procedure through process feedback is one of the main defining features of MPC.

In this framework, Vargas-Villamil (1997) propose to pose the long-term semiconductor fab scheduling as the solution of a multiobjective constrained optimisation problem:

$$\min_{\Delta u_k,\ldots,\Delta u_{k+m}} J \tag{7}$$

where J can be taken as an l_1-function due to the nature of the problem and Δu_j is a control increment. Considering maximising throughput, keeping inventories at a required level and limiting the energy of input moves, the objective function can be taken as:

$$J = -\overbrace{\sum_{l=1}^{P} \left\| \Gamma^{outs} Y_{k+1/k}^{outs} \right\|_1}^{max.\,prod.} + \overbrace{\sum_{l=1}^{P} \left\| \Gamma^{Y} \left(Y_{k+1/k} - R_{k+1/k} \right) \right\|_1}^{invent.\,set} + \overbrace{\sum_{l=1}^{M} \left\| \Gamma^{\Delta U} \Delta U_k \right\|_1}^{move\,supp.} \tag{8}$$

subject to:

$$\hat{I}U_k + I_L \Delta U_{k+1/k} \geq 0 \tag{9}$$

$$C_L I_L \Delta U_k \leq C_{Av,k+1} - C_L \hat{I} u_{k-1} \tag{10}$$

$$\begin{bmatrix} S_u \Delta U_k \\ -S_u \Delta U_k \end{bmatrix} \leq \begin{bmatrix} -S_x \Delta x_k - \hat{I} y_k - S_d \Delta D_k + Y^{\max}_{k+1/k} \\ S_x \Delta x_k + \hat{I} y_k + S_d \Delta D_k - Y^{\min}_{k+1/k} \end{bmatrix} \tag{11}$$

$$\Gamma^Y = diag\left[\lambda_1^Y \ldots \lambda_P^Y\right] \tag{12}$$

$$\Gamma^{\Delta U} = diag\left[\lambda_1^{\Delta U} \ldots \lambda_M^{\Delta U}\right] \tag{13}$$

Eq. 9 corresponds to the input or product target sign constraints, Eq. 10 is the reentrant constraint and Eq. 11 describes the inventory level constraint. P is the output prediction horizon and M is the number of control moves for prediction. Γ^Y and $\Gamma^{\Delta U}$ are matrices of output and input weights, respectively. $Y_{k+1/k}$ is the predicted output trajectory and ΔU_k contains the input control moves.

2.2 LP-MPC formulation

As the objective function is l_1-norm, the MPC problem can be transformed to an LP through a change of variables (Dave et al., 1992; Vargas-Villamil, 1997). However, for the multiobjective function in Eq. 8, the problem in the new variables gives rise to a set of constraints that becomes ill-conditioned and/or overly stringent if a desired rate for the throughput volume is specified while inventories are kept at a given set-point. This leads to numerical problems that in many cases prevent the LP from obtaining a feasible solution even if online constraint scaling is added when necessary. Note that, there will be a limit of achievable production for a given level of inventories. Under the above specifications for inventories and outputs, we consider the objective function be reduced to:

$$J = \overbrace{\sum_{l=1}^{P} \left\| \Gamma^Y \left(Y_{k+1/k} - R_{k+1/k} \right) \right\|_1}^{inv.\ and\ outs\ set\ track.} + \overbrace{\sum_{l=1}^{M} \left\| \Gamma^{\Delta U} \Delta U_k \right\|_1}^{move\ supp.} \tag{14}$$

The variable sustitution used by Vargas-Villamil *et al.* (1997) for the two terms remaining in the objective still applies. Then, taking:

$$-v_{k+1/k} \leq \rho_{k+1/k} \leq v_{k+1/k} \tag{15}$$

$$-\mu \leq \Gamma^{\Delta U} \Delta U_k \leq \mu \tag{16}$$

with:

$$v, \mu \geq 0,$$

where the residual vector ρ is given by:

$$\rho_{k+1/k} = \Gamma^Y (Y_{k+1/k} - R_{k+1}), \tag{17}$$

the problem in the new variables results in:

$$\min_{v,\mu,\Delta U_k} (v+\mu) \tag{18}$$

subject to:

$$\begin{bmatrix} -I & \Gamma^Y S_u \\ -I & -\Gamma^Y S_u \end{bmatrix} \begin{bmatrix} v \\ \Delta U_k \end{bmatrix} \leq \begin{bmatrix} -\Gamma^Y \varepsilon_{k+1/k} \\ \Gamma^Y \varepsilon_{k+1/k} \end{bmatrix} \tag{19}$$

$$\begin{bmatrix} -I & \Gamma^{\Delta U} \\ -I & \Gamma^{\Delta U} \end{bmatrix} \begin{bmatrix} \mu \\ \Delta U_k \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{20}$$

$$\begin{bmatrix} C_L I_L \\ -I_L \\ -S_u \\ S_u \end{bmatrix} \Delta U_k \leq \begin{bmatrix} C_{Av,k+1} - C_L \hat{I} u_{k-1} \\ \hat{I} u_{k-1} \\ S_x \Delta x_k + \hat{I} y_k + S_d \Delta D_k \\ -S_x \Delta x_k - \hat{I} y_k - S_d \Delta D_k + Y^{\max}_{k+1/k} \end{bmatrix} \tag{21}$$

$$v, \mu \geq 0$$

being:

$$\varepsilon_{k+1/k} = Y_{k+1/k} - R_{k+1}. \tag{22}$$

The above LP problem has $l \times P$ less optimisation variables and a well-defined set of constraints.

3 Short-Term Scheduling

A key characteristic of the short-term scheduling problem is that the planning horizon is too short for a given pattern to be established. Therefore, complete flexibility in the utilisation of resources is desirable, subject to the various technical and contractual constraints under which the plant operates.

The problem is defined by the product recipes, the plant capacities and functionalites, the utility and storage availabilities and the production objectives over a given horizon. Its solution involves the determination of detailed utilisation profiles for all resources over the same horizon, as well as the calculation of material flows in the plant.

gBSS is a multipurpose plant optimisation package developed at Imperial College, based on a STN (Fig. 2) description of the process and a general characterisation of resources. Considerable flexibility is allowed in the utilisation of processing equipment, storage capacity, manpower and utilities. Other issues, such as those related to plant operability and safety aspects, can also be taken into account whenever necessary. It includes a set of rigorous mathematical programming formulations coupled with efficient solutions algorithms.

The definition of the scheduling problem is given in three categories of information files: process recipe, plant resources and problem-specific information (Papageorgiou *et al.*, 1992). A context-specific interface manages the data in terms that are familiar to the user (Shah *et al.*, 1995). From these data files, gBSS automatically builds a mathematical formulation of the scheduling problem as a MILP, which is solved by a modified branch and bound algorithm (Shah *et al.*, 1992). The results of the optimisation are translated back into engineering terms and are displayed in a variety of informative views.

4 Case Study

Let us consider the five-machine six steps semiconductor manufacturing process presented in Vargas-Villamil (1998) and introduced in Section 2. For the purpose of comparison, we assume the same specifications.

The nominal rates in matrix B (Eq. 3) are given by: $R_1 = 7.5$, $R_2 = 12$, $R_3 = 9.6$, $R_4 = 9$, $R_5 = 6.8$, $R_6 = 24[lots/shift]$. The maximun and minimum capacity of inventories are $y^{max} = [10\,10\,10\,10\,10\,10\,\infty]^T$ $[lots]$ and $y^{min} = \mathbf{0}$. The permutation matrix in Eq. 10 is:

$$C_L = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

with the duration of shifts constrained by: $C_{Av}^{max} = [2\ 2\ 1]^T$ and $C_{Av}^{min} = \mathbf{0}[hr]$.

The boundary of performance using different initial conditions and output targets for constant work-in process inventory targets can be obtained through a Pareto curve. This establishes a trade-off between production target and throughput time. In order to avoid the wafers to be exposed to aerial contamination, inventories should be kept low, though greater than zero.

Suppose that the average initial conditions are $x(0) = [2\ 1\ 10\ 10\ 8\ 4\ 0]$. The desired inventory level is $R = [1\ 1\ 1\ 1\ 1\ 1]$, while the corresponding maximum production rate obtained from the Pareto curve is $6.38[lots/shift]$. The values of the tunning parameters for the MPC are: $P = 4$ and $M = 2$ for the prediction and input horizons, respectively. The weights are taken as: $\Gamma^Y = 1$ and $\Gamma^{\Delta U} = 1$.

The implementation of the MPC from the multiobjective in Eq. 8 was not suitable from the optimisation point of view. The resulting set of constraints after the variable transformation becomes ill-conditioned and it is not always possible to overcome the situation including on-line constraint scaling. Moreover, near the new set-point the contraints are overly stringent and the LP fails to obtain a feasible solution. No better results are got for different tunning parameters of the MPC.

The above could be inferred from the problem specifications, observing that the maximum throughput volume for a given condition has been obtained before. Then, there is no need to include the maximisation of outs in Eq. 8. Dropping this first term gives rise to the objective in Eq. 14. As long as the MPC is designed and tunned to steer the process from the initial conditions to the new set-point, the goal of maximising throughput volume while keeping inventories at set-point will be accomplished. This is shown in Fig. 3, where no plant/model mismatch is considered.

Let us include plant/model mismatch and suppose that the real rates in matrix B are: $R_1 = 7.2$, $R_2 = 12$, $R_3 = 12$, $R_4 = 12$, $R_5 = 7.2$, $R_6 = 12[lots/shift]$. New measures from the real process are available to be feed back into the MPC each shift. Keeping the controller parameters invariant with respect to the nominal case, it is observed from Fig. 4 that the starts released into the fab and the settling times have not changed significantly. The throughput volume is close to the maximum while the inventories have an offset with respect to the set-point.

The starts obtained from the MPC are fed into the short-term scheduler where the process is represented by a STN. The desired work-in inventories were specified in gBSS as 'targets'. The evolution of the inventories and the throughput are shown in Fig. 5. It is seen that the short-term scheduler can follow the policy provided by the MPC keeping the inventories in the vicinity of the target imposed. Moreover, the production level predicted by the long-term campaign planning

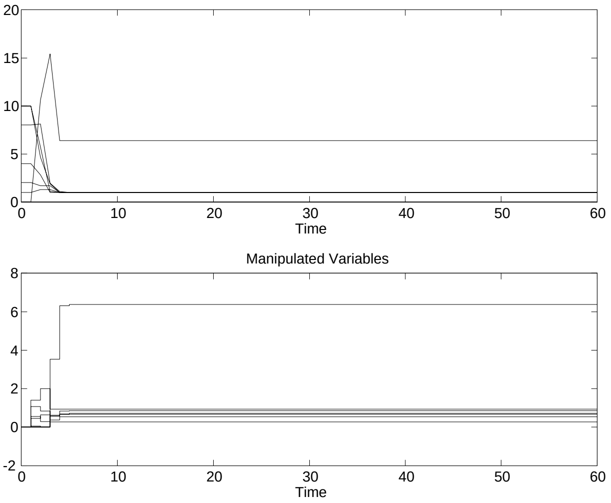

Figure 3: MPC - set-point tracking for max. throughput, open-loop

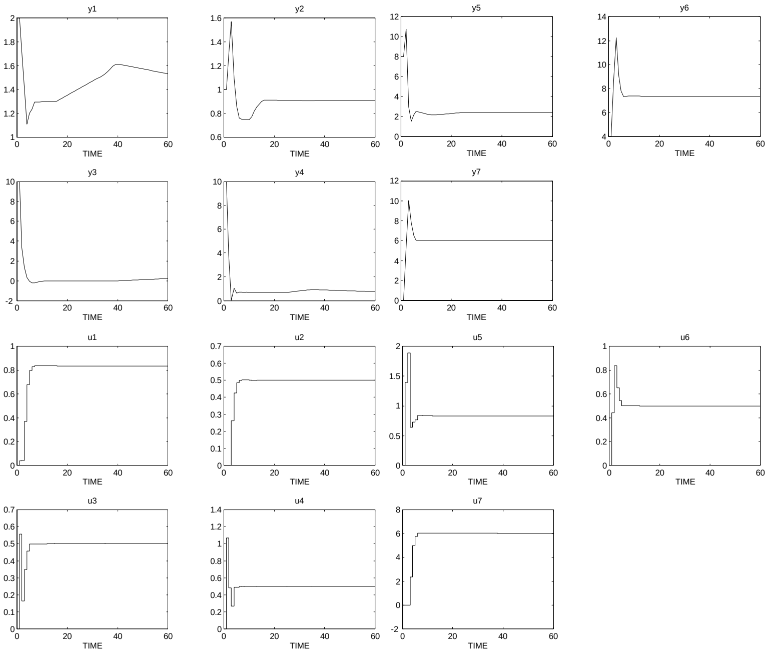

Figure 4: MPC - set-point tracking for max. throughput, closed-loop

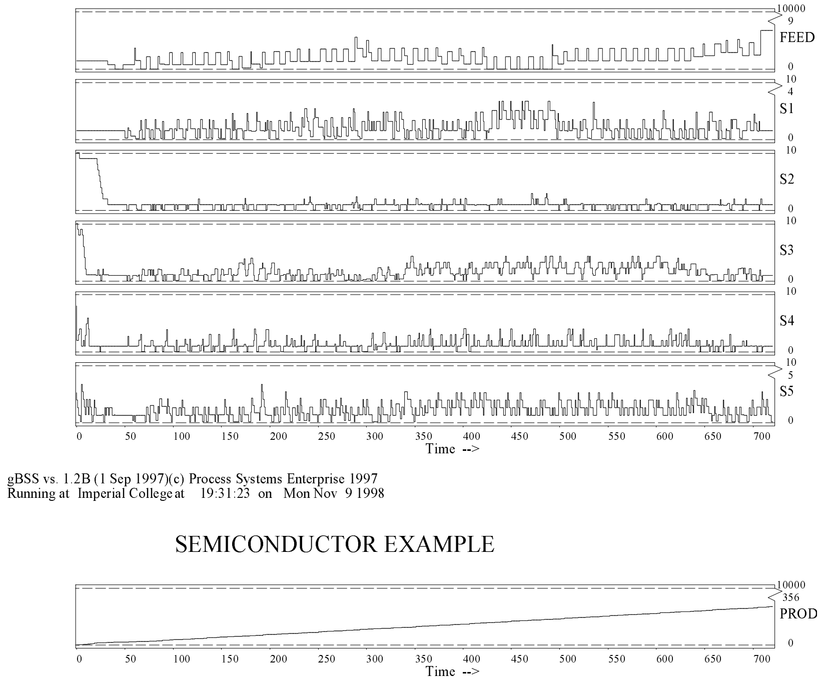

Figure 5: gBSS - inventories and throughput

is obtained in the specified horizon. The Gantt chart for the detailed operation of the fab during 2 shifts is presented in Fig. 6.

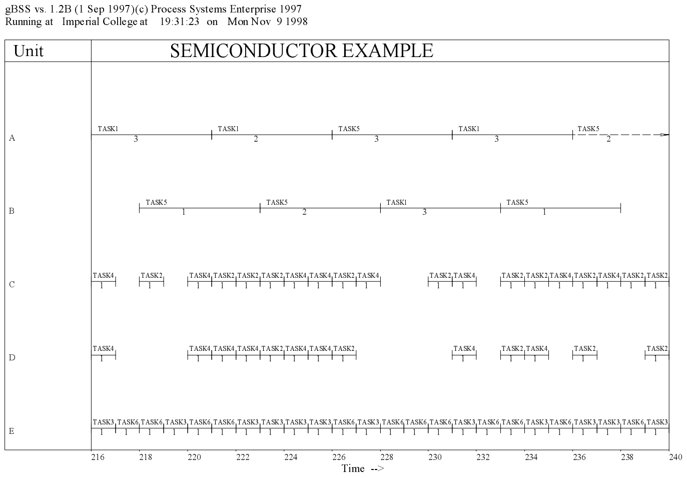

Figure 6: gBSS - detailed operation for the 5 machines, day 9th.

5 Conclusions

A proposal for long-term planning and short-term scheduling for reentrant line processes is discussed. The properties of the two different horizon planning/scheduling representations of the process in closed-loop operation are shown. As a result, detailed operation of a semiconductor manufacturing processes using gBSS was obtained, following the long-term references imposed by a MPC. The integrated software package deals with the discrete-event system dynamics through an interfase that shields the user from the complex mathematics involved in the model. Due to the optimisation carried out into gBSS, there is no need to translate the utilisation targets from MPC.

References

Dave, P., D. Willig, G. Kudva, J. Pekny, and F. Doyle (1997). "Lp methods in mpc of large-scale systems: Application to paper-machine cd control," *AIChE J.*, **43**, no. 4, pp. 1016–1031.

El Adl, A. R., M.K. and K. Tsakalis (1996). "Hierarchical modeling and control of reentrant semiconductor manufacturing facilities," in *35th. Conference on Decision and Control*, Kobe, Japan.

Garcia, C., D. Prett, and M. Morari (1989). "Model predictive control: Theory and practice - a survey," *Automatica*, **25**, no. 3, pp. 335–348.

Henson, M. and D. Seborg (1997). *Nonlinear Process Control*, Prentice Hall, New Jersey.

Kondili, E., C. C. Pantelides, and R. Sargent (1993). "A general algorithm for short-term scheduling of batch operations," *Computers and Chemical Engng.*, **17**, no. 2, pp. 211–227.

Papageorgiou, L., N. Shah, and C. C. Pantelides (1992). "A software system for optimal planning and scheduling of general batch operations," in *Advances in Process Control III*, IChemE, Rugby, pp. 161–170.

Pierce, D. and M. J. Realff (1996). "Process synthesis and design for multi-chip module fabrication," *Computers and Chemical Engng.*, **20**, no. Suppl., pp. S1307–S1315.

Shah, N., K. Kuriyan, L. Liberis, C. C. Pantelides, L. Papageorgiou, and P. Riminucci (1995). "User interfaces for mathematical programming based multipurpose plant optimisation systems," *Computers and Chemical Engng.*, **19**, no. Suppl., pp. S765–S772.

Shah, N., C. C. Pantelides, and R. Sargent (1992). "A general algorithm for short-term scheduling of batch operations: Part ii - computational issues," *Computers and Chemical Engng.*, **17**, no. 2, pp. 229–244.

Tsakalis, K., J. J. Flores Godoy, and A. A. Rodriguez (1997). "Hierarchical modeling and control for reentrant semiconductor fabrication lines: A mini-fab benchmark," in *6th. IEEE Int. Conference on Emerging Technologies and Factory Automation*, ETFA'97, Los Angeles, CA.

Vargas Villamil, F. and D. E. Rivera (1997). "Scheduling of reentrant manufacturing lines using a model predictive control (mpc) approach," in *American Control Conference*, ACC'97, vol. 3, pp. 1919–1923.

Vargas Villamil, F. and D. E. Rivera (1998). "Scheduling of semiconductor manufacturing facilities using a model predictive control approach," Tech. rep., Control Systs. Laboratory, University of Arizona, Tempe, Arizona.