Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

AVAILABLE-TO-PROMISE SYSTEMS IN THE SEMICONDUCTOR INDUSTRY: A REVIEW OF CONTRIBUTIONS AND A PRELIMINARY EXPERIMENT

Jose M Framinan
Paz Perez-Gonzalez

Industrial Management - School of Engineering
University of Seville
Ave. Descubrimientos s/n
Seville 41092, SPAIN

ABSTRACT

This paper focuses on Available-To-Promise (ATP) systems in the semiconductor industry. These systems have been successfully applied in a number of sectors, although it is often mentioned that their advantages increase with the ability of obtaining accurate forecasts, and with the possibility of identifying a relatively large number of different customers or customer classes. These conditions do not necessarily fulfil in the semiconductor industry, therefore it is interesting to analyse the few case studies of these systems that have been presented in the literature. A preliminary experiment is carried out using a foundry plant data to investigate the influence of the forecast accuracy and forecast bias in the performance of these systems. The results highlight the problems caused by the lack of homogeneity in the forecast, and the distortion introduced by customers 'inflating' their projected demand in order to ensure a higher share of the orders.

1 INTRODUCTION

Many manufacturing firms use the so-called Available-To-Promise (ATP) systems (see e.g. (Kilger and Schneeweiss 2008)) to match supply and demand while maximizing profitability. The key idea behind this type of systems is to allocate the projected stock and planned production quantities to different customers or customer segments (also called *customer classes*) to perform a real-time promising of incoming orders. The potential benefits of ATP systems become particularly relevant if customer heterogeneity is high enough, and reliable information about customer demand is available (Meyr 2009). However, in many cases within the semiconductor industry, the number of customer classes is limited (i.e. most customers are equally important), and it is difficult to acquire reliable information about customer demand (e.g. customers often report higher demand to sales managers to ensure sufficient capacity, see e.g. (Chiang and Hsu 2014)). Given these conditions, it is interesting to analyse the scope of application of these systems for semiconductor manufacturing.

In this paper, we first analyse the main decisions involving ATP functionalities. Next, we review the application of ATP in the semiconductor industry, as well as in related sectors. Using this information, we discuss the main features of the sector affecting ATP-related decisions. It turns out that the asymmetry of the customers with respect to their priorities and/or forecasting may be a major issue when implementing these systems in the semiconductor industry. We use a set of data from a foundry plant to assess the effect of forecasting asymmetry in planning allocation policies.

The remainder of the paper is as follows: in Section 2 we introduce the ATP systems in order to provide a framework to classify existing descriptions of ATP systems in the semiconductor and related industries, which we carry out in Section 3. The review reveals the potential importance of forecast accuracy and forecast bias in the performance of ATP systems, so a series of preliminary experiments are conducted.

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

The description of the experiments is presented in Section 4 whereas the results are discussed in Section 5. Finally, the conclusions of the work are presented in Section 6.

2 ATP SYSTEMS

Although there is no clear definition of an ATP, generally speaking, the following decision aspects aspects may be involved (Pibernik 2005, Framinan and Leisten 2010):

- Allocation Planning. Allocation planning deals with assigning the projected supply of a product (or technology) to a specific customer or customer class, therefore effectively dividing the capacity between the **committed ATP quotas**, and the **uncommitted ATP capacity**. The basis for this allocation decision problem is the information from Demand Planning regarding the projected demand for each customer and product on each period (usually originated from sales forecasts, preorders, or both), and from Capacity Planning regarding the projected capacity for each product. In a multi-factory setting, allocation planning can be done in an integrated manner (committed ATP quotas are decided for each product and for each plant), or in a two-step procedure: the demand is first allocated to each plant (possibly taking into account the different margins), and then each plant calculates its own committed ATP quotas. We will refer to this last option as a *decentralised allocation planning* (see e.g. (Vogel and Meyr 2013)), which obviously yields suboptimal solutions with respect to the former option. However, since some companies maintain certain degree of autonomy in the management of their plants, this option is sometimes employed.
  Since allocation planning requires the projected supply as an input data, this data can be provided at different levels of granularity. More specifically:
  - Capacity planning level, i.e. the projected products is provided at the aggregation level given by capacity planning, usually at product level in a month-interval.
  - Shop floor level, i.e. the projected products are provided at shop floor level, usually in a daily basis.

  Most references use the data at capacity planning level, however (Jeong et al. 2002) describe a case where input data are provided at shop floor level.

  Furthermore, different allocation planning rules can be employed, including:
  - No allocation. That means that all projected supply is available for the order promising stage. If all customers are equally important, then this is the allocation planning rule with the highest flexibility.
  - Customer-class based. That means that the profit (cost) for each customer or class of customers is different, therefore allocation can be performed to book capacity for high-profit customers that may be late. Note that this would also include the case where the profit (cost) is different depending on the sales channel. Hard pegging the projected products to classes of customers or to specific customers may imply that, at the end of the allocation planning procedure, there is no uncommitted ATP capacity, particularly if the demand forecasts exceed the available capacity. This raises the issue of establishing caps on the capacity to be allocated, thus always leaving unallocated a fraction of the available capacity. The allocation to customer classes based on profit is also named capacity nesting (Gossinger and Kalkowski 2015).
- ATP consumption (order promising). As firm orders from customers arrive, they have to be matched with their committed ATP quotas. If the order cannot be fulfilled from the committed ATP quotas, they may be fulfilled from the uncommitted ATP capacity. Regarding ATP consumption, usually two *modes* can be observed:
  - Real-time ATP. In a real-time ATP, the order is promised immediately after the customer places the order. In order words, the ATP consumption is activated after receiving each order.

  Real-time ATP decisions have been also labelled in the scheduling literature as job-insertion problem (Roundy et al. 2005).
  - Batch ATP. During a batching interval, orders are first collected and then processed together. In this manner, the orders in the batching interval can be optimally allocated.

  Although most related literature makes a clear distinction between these two aspects, some systems combine both: in (Ball, Chen, and Zhao 2004) a two-step approach is described which is employed consisting first of giving a soft (and coarse) commitment to the customer in real time, and after a few days giving a hard (and more accurate) commitment after collecting several orders and allocating them using a batch ATP.
- ATP reallocation. As the committed ATP quota may not be realised by the corresponding customer orders, a reallocation mechanism may be devised to avoid that this committed capacity becomes unused. Regarding ATP reallocation, a number of policies may be adopted, including:
  - No reallocation: unused committed ATP quotas are not released.
  - Cut-off list: In (Chiang and Hsu 2014) a mechanism is described to urge the confirmation of the orders by the customers as their booking approaches the date in which their committed ATP cannot be used by any other customer.

The above summary has served us to highlight the main decisions/policies in ATP systems. In the next section, we will use it to describe existing ATP in the semiconductor and related industries.

3 CASE STUDIES IN SEMICONDUCTOR AND RELATED INDUSTRIES

Some characteristics (although not specific to) of the semiconductor industry:

- Utilization rate of capacity is very high, sometimes exceeding 90% (Chiang and Wu 2011, Semiconductor 2010). This feature is shared with a number of sectors, including tool machine manufacturing, or glass contained manufacturing (Sridharan 1998).
- Forecasting is extremely complicated, incurring in considerable uncertainty (Chiang and Hsu 2014), so a deviation from the input provided by Demand Planning has to be seen as unavoidable. This stresses the need of proper ATP re-allocation mechanisms.
- Semiconductor foundry customers often report higher demand to sales managers to ensure sufficient capacity (Chiang and Hsu 2014), which seems to be a by-product of the relative lack of capacity in the sector, although it is not exclusive of the semiconductor industry (Lee, Padmanabhan, and Whang 2004).
- Many companies have a relatively low number of customers, so few customers represent most sales of the company.

A number of papers have been published regarding the application of ATP systems in semiconductor and other related industries. (Yang et al. 2010) discuss the advantages and disadvantages of using ATP systems in the semiconductor industry as opposed as a First Come First Served (FCFS) policy for order fulfilment. An early experience in the TFT-LCD industry is (Jeong et al. 2002). In their system, the granularity of the input is very low (shop floor level), so the available capacity is determined by the production schedule and not by capacity planning. This allows having detailed capacity data across the supply chain and to allocate orders according to a real-time mode, which is the problem addressed in their paper. Also in the TFT-LCD industry, (Lin and Chen 2005) describe a case of an ATP system in which preferences for products and plants can be expressed in the allocation planning procedure, thus not only mazimizing the revenues but the preferences of important customers. ATP consumption is not discussed in their paper. (Tsai and Wang 2009) describe an ATP system in the TFT-LCD industry where the allocation is carried out in a decentralised manner. Allocation planning is thus carried out in two phases, and then a batch-ATP mode is run. Another implementation is described in (Lin, Hong, Wu, and Wang 2010), where the trade-off between the batch size and the benefits of the ATP system are discussed. (Han, Dong, and Liu 2014) describe an allocation planning policy if substitute (upgrade) products can be used to satisfy demand not previously allocated.

Finally, an order admission mechanism in the semiconductor industry is discussed in (Chiang and Hsu 2014). The system described uses an allocation planning procedure based on a linear programming model that seeks to maximise the revenue. Experiments on ATP usage are carried out in real-time and batch modes, showing little differences regarding both options. A re-allocation mechanism based on a cut-off list is discussed, showing the advantages of ATP reallocation, particularly when demand forecasts are not accurate.

As a summary, few papers address ATP systems in the semiconductor or related industries. Most of these papers focuses on stressing the potential advantages of ATP, particularly for customer segmentation. In all the cases reviewed, input data for allocation planning are obtained from forecast, and not from firm orders. Only in the paper by (Chiang and Wu 2011), a discrepancy between forecast and firm orders is explicitly addressed. However, no mechanisms to handle this discrepancy –such as setting some limits in the quotas that are assigned to specific customers– are discussed. Therefore, it seems interesting to address these aspects in order to better assess how the forecast may influence the performance of ATP systems.

4 EXPERIMENTAL EVALUATION

In this section, we conduct a series of preliminary experiments to better understand the trade-off between the forecast accuracy and bias, and the allocation planning rules. Two performance indicators will be used in these experiments:

- Fraction of unused capacity (*UC*), i.e. the fraction of capacity that could not be used as it was been allocated to specific customers who do not issued orders according to their forecasts/pre-orders.
- Service Level (*SL*), i.e. the fraction of orders that are fulfilled. Equivalently, the fraction of unattended orders $(1-SL)$ can be measured.

The following controllable factors have been considered in the experiments:

- **Forecast Accuracy**. Clearly, forecast accuracy influences the unused capacity, as extremely accurate forecasts would produce committed ATP quotas that would eventually be fulfilled by customer orders. An widely used indicator to compute forecasting accuracy is the Mean Average Percentage Deviation (MAPE), measured as follows:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}|\frac{d_t-\hat{d_t}}{d_t}| \tag{1}$$

  where d_t is the actual demand in period t, and $\hat{d_t}$ is the forecasted demand in period t. According to the interpretation by (Lewis 1982) of MAPE, in the experiments we use the following values:
  - 15% for all customers/class, which in (Lewis 1982) interpretation corresponds to a good forecasting. This factor is denoted as `ACCURATE`.
  - 30% for all customers/class, which in (Lewis 1982) interpretation corresponds to a good forecasting. This factor is denoted as `ACCEPTABLE`.
  - 70% for all customers/class, which corresponds to an inaccurate forecasting. This factor is denoted as `INACCURATE`.
  - Non homogeneous forecasting. In most works, it is considered that the forecasting is homogeneous, i.e. the forecasting error is similar across customers. In our experiments, we would like to test how different degrees of accuracy in forecasting affect to the system. Therefore, we also consider a case where the forecast is good for a class of customers (15% MAPE), reasonable

for another class of customers (30% MAPE), and inaccurate (70% MAPE) for another class of customers. This factor is denoted as `NON-HOMOGENEOUS`.

Note that other indicators to compute forecasting accuracy, such as SMAPE, are also widely used in the semiconductor industry (see e.g. (Habla et al. 2007)). However, we are not aware of a similar interpretation for SMAPE such as the one in (Lewis 1982).

- **Forecast Bias**. In addition to measure different levels of forecast accuracy, the bias in the forecasting can be also measured. By bias we mean that the actual orders are consistenly underestimated or overestimated for a customer, i.e. the average realization of orders does not coincide with the forecast for this customer. To the best of our knowledge, this aspect has not been investigated in the related literature and we believe that it may be particularly relevant for the semiconductor industry, as it has been discussed previously that some customers may report higher demands to sales managers. Therefore, we consider the following level of bias:
  - No bias, i.e. that the average number of orders for each customer class are centered around their MAPE errors. This factor is denoted as `NO-BIAS`
  - Positive bias for a class of customers. More specifically, one class of customers reports demands which are, on average, 20% higher and 50% higher to those finally realised. These factors are denoted `SMALL-OVER` and `BIG-OVER`, respectively. The scenario of a positive bias may serve to model the case of a customer/class of customers who may ‘inflate’ his demands in order to make sure that his orders are fulfilled.
  - Negative bias for a class of customers. More specifically, for one class of customers, their demands are underestimated by 20%, and 50%. These factors are denoted `SMALL-UNDER` and `BIG-UNDER`. This scenario serves to explore the consequences of underestimating the demand due to e.g. performing conservative forecasts for customers who inflated their demand in the past.
- **Cap on the allocated capacity**. Since a factor to accommodate the uncertainty regarding the forecast is the α level (i.e. the fraction of capacity which is committed), we consider the following values of $\alpha := (0.5, 0.8, 1.0)$.

The experiments carried out consist on replicating the decisions in a typical ATP system. More specifically, the ATP system under study is composed of two stages:

1. Stage 1: Allocation Planning. On the basis of the forecasts from Demand Planning, together with the capacities obtained from Capacity Planning, the committed ATP quotas are determined.
2. Stage 2: ATP consumption. Once the committed ATP quotas and uncommitted ATP capacity are determined, the customer orders start to arrive and are fulfilled in a real-time mode (i.e. FCFS).

The allocation planning stage is carried out using a Linear Programming model that maximises the expected profit. This model is described in greater detail in (Chiang and Hsu 2014). More specifically, the LP model employed uses variables X_{cfgt} to determine the committed ATP quota for customer c ($c = 1,\dots,C$) regarding the production in factory f ($f = 1,\dots,F$) of product g ($g = 1,\dots,G$) in period t ($t = 1,\dots,T$). Assuming that M_{cfgt} represents the margin obtained per unit of product g produced for customer c in factory f in period t, the objective function may be stated as follows:

$$\max \sum_{c=1}^{I} \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{t=1}^{T} M_{cfgt} X_{cfgt} \tag{2}$$

subject to the following constraints:

$$\sum_{c=1}^{C} X_{cfgt} \leq \alpha \cdot CAP_{fgt} \quad f = 1,\dots,F; g = 1,\dots,G; t = 1,\dots,T \tag{3}$$

$$X_{cfgt} \leq D_{cfgt} \quad c = 1,\ldots,C; f = 1,\ldots,F; g = 1,\ldots,G; t = 1,\ldots,T \tag{4}$$

$$X_{cfgt} \geq QTY_{cfgt} \quad c = 1,\ldots,C; f = 1,\ldots,F; g = 1,\ldots,G; t = 1,\ldots,T \tag{5}$$

$$X_{cfgt} \geq 0 \quad c = 1,\ldots,C; f = 1,\ldots,F; g = 1,\ldots,G; t = 1,\ldots,T \tag{6}$$

Constraints (3) ensure that the committed ATP quotas of a given class of product g produced in factory f allocated to all customers in time period t do not exceed a fraction α of the capacity of the factory for the product in that period (CAP_{fgt}). In many allocation planning models – such as in (Chiang and Hsu 2014) –, $\alpha = 1$, indicating that all capacity may be 'hard-pegged' to specific customers. However, in our case, this limit on the allocated capacity will be a factor in the experiments, as discussed before.

Constraints (4) ensure that the committed ATP quotas do not exceed D_{cfgt} the demand forecast of product g for customer c in period t to be fulfilled from factory f. Constraints (5) ensure that the firm orders existing at the time of running the model are allocated. More specifically, QTY_{cfgt} the required quantity (firm orders) by customer c of product g in period t to be manufactured in factory f has to be fulfilled.

The output of the previous model determines the committed ATP quotas (variables X_{cfgt}), while $UATP_{fgt}$ is the uncommitted ATP capacity of factory f for product g in period t, and can be obtained according to the following expression:

$$UATP_{fgt} = CAP_{fgt} - \sum_{c=1}^{C} X_{cfgt} \tag{7}$$

Once X_{cfgt} and $UATP_{fgt}$ have been determined, the corresponding supply is allocated either to specific customers/customer classes, or to the so-called 'central pool' where it can be assigned in Stage 2 to specific customer firm orders on a FCFS basis. Note that we perform a real-time ATP since, despite batch ATP is prevalent in many industries, it is considered a drawback in the semiconductor sector, as it does not provide real-time decision support for sales managers (Chiang and Wu 2011).

Therefore, we take the committed ATP quotas and the uncommitted ATP capacity, and simulate the arrival of the firm customer orders. To do so, we first generate TO_{cgt} the total quantity of product g to be requested by customer class c in period t (total firm orders of the product placed by each customer class to be fulfilled in period t). This total amount may or not be in line with the demand forecast for each customer and product on each period. We will use the factors FA_c to denote the Forecast Accuracy for customer c, and FB_c to denote the Forecast Bias of customer c (discussed above) to control the extent to which the actual total demand matches the forecast used in the allocation planning stage. Therefore, TO_{cgt} are drawn according to a uniform distribution in the following interval:

$$[D_{cgt} \cdot FB_c(1 - FA_c), D_{cgt} \cdot FB_c(1 - FA_c)] \tag{8}$$

where the values of FA_c and FB_c correspond to the different scenarios considered above:

- Regarding FA_c:
  - $FA_c = 0.15\ \forall c$ corresponds to scenario `ACCURATE`
  - $FA_c = 0.30\ \forall c$ corresponds to scenario `ACCEPTABLE`
  - $FA_c = 0.70\ \forall c$ corresponds to scenario `INACCURATE`
  - $FA_1 = 0.15, FA_2 = 0.3, FA_3 = 0.7$ corresponds to scenario `NON-HOMOGENEOUS`
- Regarding FB_c:
  - $FB_c = 1.0\ \forall c$ corresponds to scenario `NO-BIAS`
  - $FA_1 = 1.0, FA_2 = 1.0, FA_3 = 0.8$ corresponds to scenario `SMALL-OVER`

- $FA_1 = 1.0, FA_2 = 1.0, FA_3 = 0.5$ corresponds to scenario `BIG-OVER`
- $FA_1 = 1.0, FA_2 = 1.0, FA_3 = 1.2$ corresponds to scenario `SMALL-UNDER`
- $FA_1 = 1.0, FA_2 = 1.0, FA_3 = 1.5$ corresponds to scenario `BIG-UNDER`

Once TO_{cgt} have been generated, we conduct a simulation to replicate the arrival of the specific order O_{icgt}, which denotes order i consisting of the request of product g for customer c in period t. Obviously $TO_{cgt} = \sum_i^I O_{icgt}$. More specifically, we generate a number of orders which amounts, on average, to one per day per customer per product. O_{icgt} are drawn from a uniform distribution in the interval $[250, 800]$ units, which is based on the data provided by (Chiang and Hsu 2014). The customer and product are chosen at random, and the period t for which the order is requested is generated with a probability which is higher for periods closer to the simulation clock. This serves to model the situation where orders for the next month are more likely that for several months in advance. Once order O_{icgt} is generated, it is attempted to be fulfilled first from the committed ATP quota for this customer, product and period. If at least part of the order can be fulfilled from the committed ATP quota (i.e. $X_{cfgt} > 0$), then X_{cfgt} is updated accordingly, i.e. $X_{cfgt} = \max\{X_{cfgt} - O_{icgt}; 0\}$. Otherwise, the system attempts to fulfil the order (or the part of the order that cannot be fulfilled from X_{cfgt}) from $UATP_{fgt}$. If this is possible, the uncommitted ATP capacity is updated accordingly. Finally, if this is not possible –there is not enough uncommitted ATP capacity–, then part of the demand (of all the demand) cannot be fulfilled. The quantity of product in an order that can be fulfilled in each period is recorded, as well as the units that cannot be fulfilled, in order to obtain at the end of the simulation the unused capacity UC and the unattended orders $1 - SL$.

The simulation of this ATP system has been modelled with Microsoft Visual Studio using C#, calling the .NET libraries of the solver Gurobi 6.5.1. to solve the allocation planning linear programming model. Input data have been extracted from (Chiang and Hsu 2014). More specifically, the following data have been used:

- Size of the model: 3 customers, 3 products, 1 factory, and 12 time periods (months).
- Monthly forecast demand for each customer and product.
- Monthly capacity for each product.

In addition, in order not to introduce a distortion in the results due to the consideration of high-revenue customers, we assume that all customers have the same margins for all products. This is equivalent to maximise the expected capacity and therefore, to minimise expected unused capacity. Note that, since we do not consider product substitution and do not set different margins for the products, similar insights could be obtained with only one product. However, in order to use the same real data as in (Chiang and Hsu 2014), we keep their original setting.

5 RESULTS

The results have been analysed using Analysis of Variance (ANOVA) techniques. 50 simulations for each aforementioned factor have been considered. Figures 1-4 summarise the main results obtained. In view of the figures, the following comments can be done:

- Imposing limits on the capacity that can be allocated to customers/customer classes seems to be a good strategy to reduce the dependency on the accuracy of the forecast. As it can be seen in Figure 1, the unused capacity of the system is less sensitive to the accuracy of the forecast (i.e. for $\alpha = 0.5$ the curve is roughly flat). The highest the level of α, the more sensitive the system is to the accuracy of the forecast, which is a foreseeable result.
- Figure 1 also speaks for the importance of an homogeneous forecast among customers/customer classes. It is preferable to have acceptable but homogeneous forecasts than a non homogeneous one, despite the specific level of α that are employed.

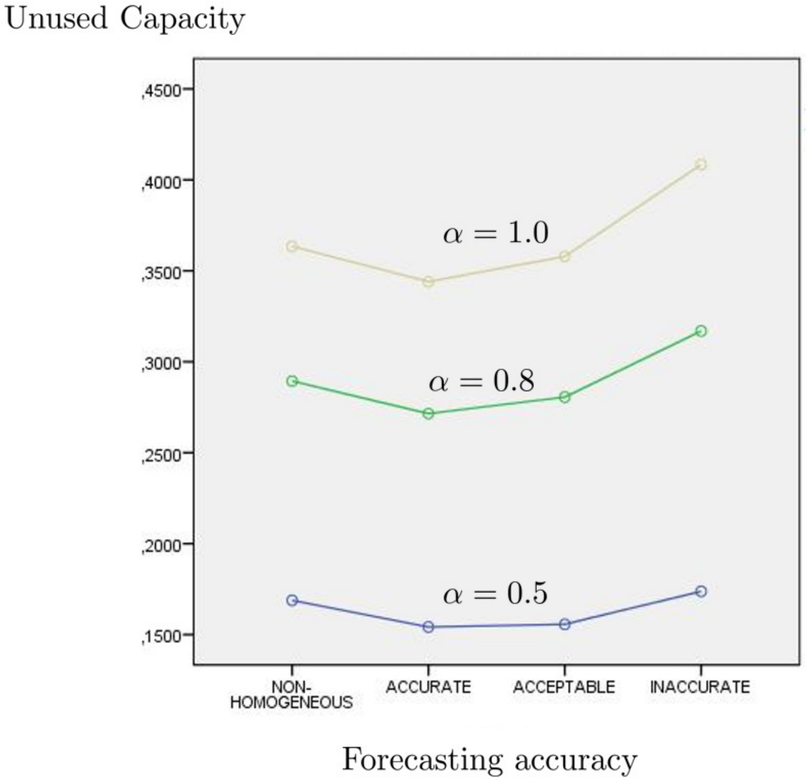

Figure 1: Unused capacity vs. forecasting accuracy.

- Despite the ability of imposing limits on the capacity allocation for handling the inaccuracy in the forecast, Figure 2 shows that this strategy cannot cope with the bias in the estimation, as the pattern is the same for different values of α.
- From Figure 2 it can be seen that forecast under estimation leads to a lower level of unused capacity, although this comes at a price of a lowest service level (see Figure 3), which is clearly an undesirable situation. In other words, there is a rather heavy penalty for underestimation in terms of unattended orders. The interpretation is that it may be rather useless to establish classes of customers to 'book' their future demand if, after all, this demand is underestimated and therefore a sizeable part of the actual orders would have to be fulfilled from the uncommitted ATP capacity.
- It can be seen from Figure 2 that one customer/customer class 'inflating' the demand to ensure a higher share of the orders causes a large increase in the unused capacity.
- Finally, as it can be seen in Figure 4 regarding the effect of forecasting accuracy in the number of unattended orders, a non homogeneous forecast is the second worst situation after an inaccurate forecast.

6 CONCLUSIONS

In this paper, we have discussed the use of ATP systems in the semiconductor and related industries. It is often mentioned that the advantages of ATP systems increase with the ability of producing accurate forecasts, and with the possibility of identify a relatively large number of different customers or customer classes. These conditions do not necessarily fulfil in the semiconductor industry, therefore it is interesting

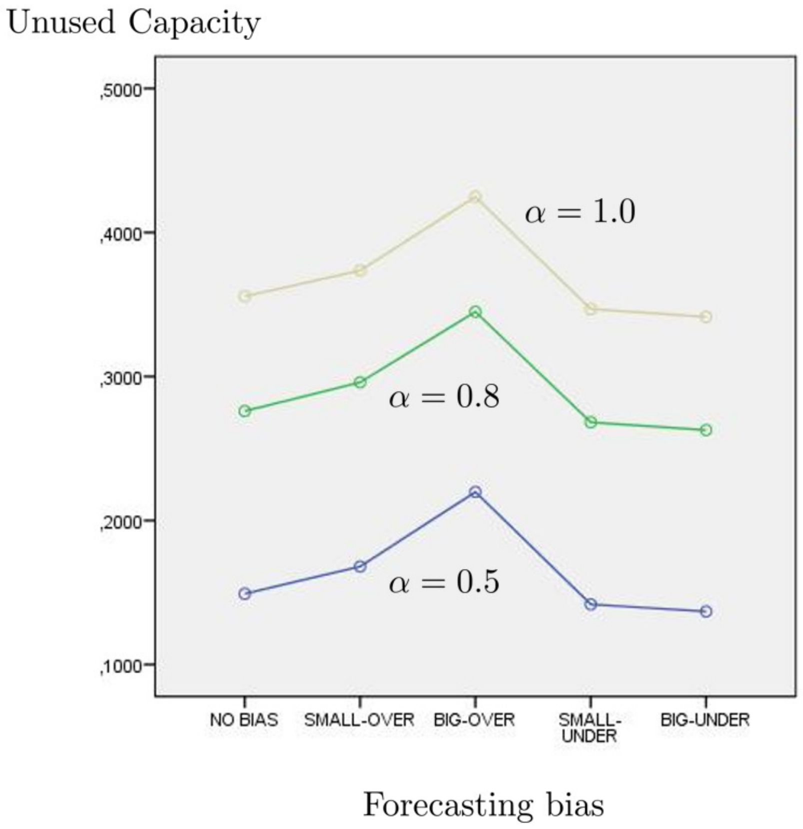

Figure 2: Effect of the forecasting bias in the unused capacity for different values of α.

to analyse the few case studies that have been presented in the literature. An experiment carried out based on real data reveals the sensitivity of ATP allocation to the accuracy and the bias of the forecasts. While imposing caps on the allocated capacity may be seen as a remedy to handle forecast inaccuracy, this strategy has little effect for handling the bias in the forecasts for certain class of customers. A particular source of problems is one customer/customer class 'inflating' the demand to ensure a higher share of the orders, which may indicate the need to rely on historical information to filter the requests for this profile of customer.

Another problem relates to lack of homogeneity of the forecasts, which may speak for the need of a centralised demand planning in order to ensure forecasting procedures that should be consistent for the different customers or customer classes.

There are a number of limitations in the experiment presented in Section 4: First, the data come from a single case, therefore it is not possible to generalise the conclusions. Second, the effect of the allocation review mechanisms has not been studied. These mechanisms, if properly implemented, can reduce the differences between the base cases and the scenarios with higher inaccuracy and bias. Third, forecast accuracy is assumed not to be dependent on the proximity of t with respect to the actual period (see equation (8)). However, it is reasonable to assume that forecast accuracy increases as t decreases. Finally, there are additional factors that can be analysed, such as the ratio of actual demand vs. capacity, or the fact that not all customers yield the same profit. However, despite its work-in-progress nature, we hope that this work serves a starting point to gather more data and assumptions related to ATP systems in the semiconductor industry, so more general testbeds can be developed.

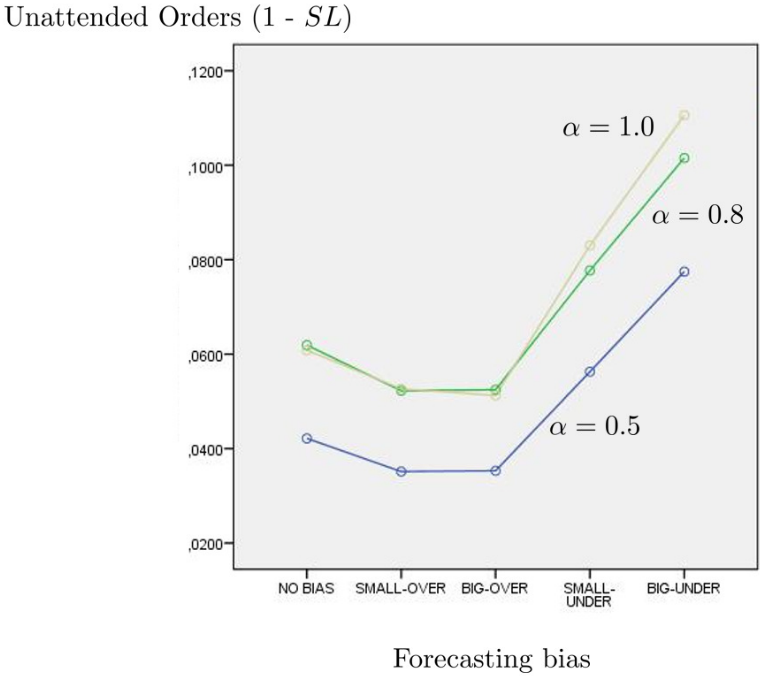

Figure 3: Effect of forecasting bias in the level of service for different values of α.

ACKNOWLEDGMENTS

The authors are sincerely grateful to the anonymous referees, who provide very valuable comments on an earlier version of the paper. This research has been funded by the Spanish Ministry of Science and Innovation, under project “ADDRESS” with reference DPI2013-44461-P/DPI.

REFERENCES

Ball, M., C.-Y. Chen, and Z.-Y. Zhao. 2004. “Available to promise”. *Handbook of Supply Chain Analysis in the EBusiness Era*:447–484.

Chiang, C., and H.-L. Hsu. 2014, nov. “An Order Fulfillment Model With Periodic Review Mechanism in Semiconductor Foundry Plants”. *IEEE Transactions on Semiconductor Manufacturing* 27 (4): 489–500.

Chiang, D. M. H., and A. W. D. Wu. 2011. “Discrete-order admission ATP model with joint effect of margin and order size in a MTO environment”. *International Journal of Production Economics* 133 (2): 761–775.

Framinan, J., and R. Leisten. 2010. “Available-to-promise (ATP) systems: A classification and framework for analysis”. *International Journal of Production Research* 48 (11): 3079–3103.

Gossinger, R., and S. Kalkowski. 2015. “Robust order promising with anticipated customer response”. *International Journal of Production Economics* 170:529–542.

Habla, C., R. Driebel, L. Monch, T. Ponsignon, and H. Ehm. 2007, Sept. “A Short-Term Forecast Method for Demand Quantities in Semiconductor Manufacturing”. In *2007 IEEE International Conference on Automation Science and Engineering*, 94–99.

Han, G., M. Dong, and S. Liu. 2014. “Yield and allocation management in a continuous make-to-stock system with demand upgrade substitution”. *International Journal of Production Economics* 156:124–131.

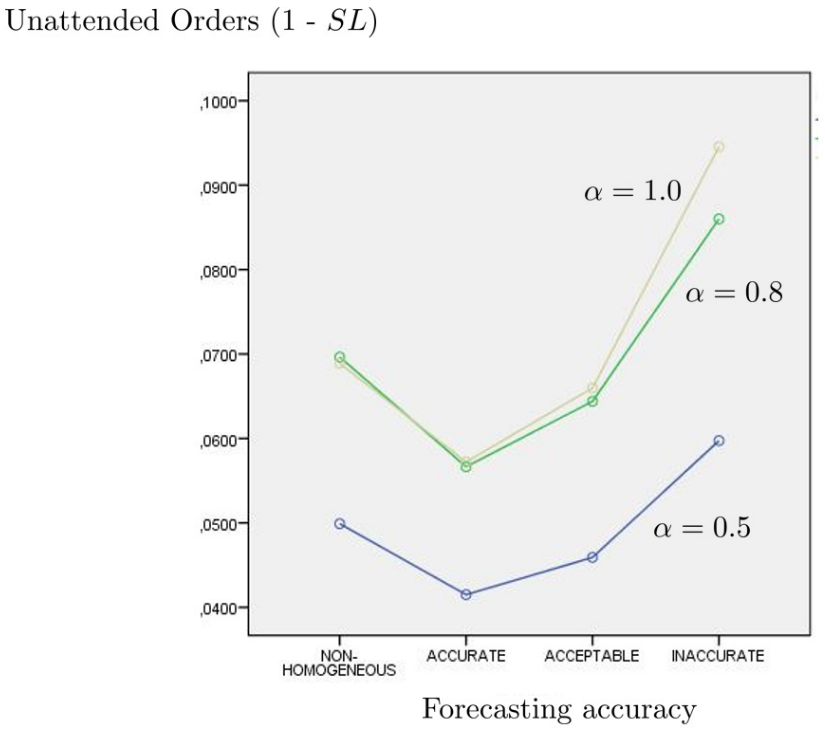

Figure 4: Effect of forecasting accuracy in the level of service for different values of α.

Jeong, B., S. B. Sim, H. S. Jeong, and S. W. Kim. 2002. "An available-to-promise system for TFT LCD manufacturing in supply chain". *Computers and Industrial Engineering* 43 (1-2): 191–212.

Kilger, C., and L. Schneeweiss. 2008. "Demand fulfillment and ATP". *Supply Chain Management and Advanced Planning: Concepts, Models, Software and Case Studies (4th Edition)*:135–148.

Lee, H., V. Padmanabhan, and S. Whang. 2004. "Information distortion in a supply chain: The bullwhip effect". *Management Science* 50 (12): 1875–1886.

Lewis, C. 1982. *Industrial and Business Forecasting Methods*. Kent: Butterworth.

Lin, J. T., and J.-H. Chen. 2005. "Enhance Order Promising With Atp Allocation Planning Considering Material and Capacity Constraints". *Journal of the Chinese Institute of Industrial Engineers* 22 (4): 282–292.

Lin, J. T., I. H. Hong, C. H. Wu, and K. S. Wang. 2010. "A model for batch available-to-promise in order fulfillment processes for TFT-LCD production chains". *Computers and Industrial Engineering* 59 (4): 720–729.

Meyr, H. 2009. "Customer segmentation, allocation planning and order promising in make-to-stock production". *Supply Chain Planning: Quantitative Decision Support and Advanced Planning Solutions*:117–144.

Pibernik, R. 2005. "Advanced available-to-promise: Classification, selected methods and requirements for operations and inventory management". *International Journal of Production Economics* 93-94 (SPEC.ISS.): 239–252.

Roundy, R., D. Chen, P. Chen, M. Çakanyildirim, M. B. Freimer, and V. Melkonian. 2005. "Capacity-driven acceptance of customer orders for a multi-stage batch manufacturing system: Models and algorithms". *IIE Transactions (Institute of Industrial Engineers)* 37 (12): 1093–1105.

Semiconductor, S. 2010. "IC Capacity Nears 100%". *Silicon Semiconductor*.

Sridharan, V. 1998. “Managing capacity in tightly constrained systems”. *International Journal of Production Economics* 56-57:601–610.

Tsai, K. M., and S. C. Wang. 2009. “Multi-site available-to-promise modeling for assemble-to-order manufacturing: An illustration on TFT-LCD manufacturing”. *International Journal of Production Economics* 117 (1): 174–184.

Vogel, S., and H. Meyr. 2013. “Decentral allocation planning in multi-stage customer hierarchies”. *European Journal of Operational Research* 246 (2): 462–470.

Yang, L., K. Chen, Z. Lin, R. Qiang, X. Huang, and R. Lin. 2010. “Order promising with capacity reserved for multi-priority orders”. In *2010 International Conference on Management and Service Science, MASS 2010*.

AUTHOR BIOGRAPHIES

JOSE M FRAMINAN is a Professor of Industrial Engineering in the School of Engineering at the University of Seville. His research interests include decision support systems and procedures for operations and supply chain management. His email address is framinan@us.es.

PAZ PEREZ-GONZALEZ is an Assistant Professor in the School of Engineering at the University of Seville. Her research interests include statistics, simulation and scheduling. Her email address is pazperez@us.es.