IT CAPACITY FORECASTING: STATISTICAL MODELLING PROCESS AND APPLICATIONS FOR SEMICONDUCTOR INDUSTRY.

M. Lutz* **. X. Boucher.*
O. Roustant*. X. Ambrosioni**

*ENSMSE-G2I, 158 cours Fauriel, 42100 Saint Etienne, France
(Tel: + 33-4-38 92-21-98; e-mail: lutz@emse.fr).
** STMicroelectronics, Crolles, France

Abstract: The objective of this paper is to present a modeling procedure, based on the use of statistical methods, to develop predictive models for Information Technology (IT) capacity planning. The goal of these models is to be able to predict IT variables’ behaviour, using initial inputs provided by high-level business explanatory variables. The modelling process consists in two main sequences, exploration and explanation, divided into modelling steps, which mix qualitative and quantitative analyses. The modelling process builds a progressive understanding of relationships among the different levels of the IT architecture. The final outcome is a statistical model, linking IT resource utilization to manufacturing activities variables. An application of the modelling process is developed in the context of the STMicroelectronics information system. As a result, two predictive models which focus on subsystems of the whole information system, link the physical servers’ daily use to manufacturing variables such as level of production, manufacturing processes deployment and fabrication throughput.

Keywords: information system, IT capacity planning, statistical analysis, semiconductor industry

1. INTRODUCTION AND RESEARCH BACKGROUND

An Information System is a critical factor when bettering modern manufacturing systems efficiency (Cardinali, 1992; Gowan and Mathieu, 1996). This is notably true within the current application field: semiconductor industry (Leachman and Hodges, 1996). Information System’s efficiency itself is based on an adequate specification of firms’ Information Technologies (IT) architectures (Earl, 1989). Consequently, IT managers must basically anticipate to provide without interruption enough IT resources to support firm’s operations plans and forecasts so that they function without interruption (Kloesterboer, 2011).

Therefore, quantitative IT capacity models are of great interest to support IT managers’ decision processes. The ITIL (Information Technology Infrastructure Library) framework recommends three main approaches for IT capacity modelling: analytical & simulation modelling and as well as trend analysis (Kloesterboer, 2011). The purpose of such models is to quantify the level of IT resources needed for a given level of workload, forecasted by IT managers on the basis of hypothetical scenarios (Jain, 1991; Gunther, 2007). However, usual IT capacity planning studies in the field of computer systems engineering are still struggling to build such models, with an operational link to high-level business or industrial variables. For instance, in a recent publication it was stated: “You may not be able to make a direct correlation between the business trends and your IT capacity, but you should at least look at whatever your organization publishes and see how it compare your component and service trends” (Kloesterboer, 2011). The purpose of our paper is to present an answer to such a limitation. Indeed, we intend to demonstrate that 1) links between IT resources and high-level industrial activities (here silicon wafer manufacture) can be established; 2) a structured modelling process can formalize rigorously such links.

This paper presents some results of PhD research in collaboration with the Crolles 300 IT Department of the company STMicroelectronics. From an industrial perspective, the objective is to build a model with good predictive features, based on the link among IT resources variables and high-level industrial variables. The impact of such model would be not only to ease communication among IT and production managers but also to compare and analyse IT capacity hypothetical scenarios, based on industrial input variables. Academically, many research papers deal with the link between semiconductor manufacturing and production equipment capacity (Geng and Jiang, 2009). However, to the best of our knowledge, few if any publications link semiconductor manufacturing system activity to the capacity of its IT architecture. More generally, it can be observed that IT capacity planning has little been explored by industrial engineering as of yet. The results presented in this paper only constitute an important initial step for much further research. The final objective will be to completely design a decision-making tool, which will encompass the use of quantitative models in a wider industrial scope. However, this decision tool is not the purpose of this paper and will be addressed in future publications. The focus of the paper is to present the modelling process used to build the quantitative models that will be required.
The rest of this article is organized as follows. Section 2 specifies the background of the research and provides a general overview of a modelling process, addressing the research context described above. This process is organized through two sequences, exploratory and explanatory. Each sequence is decomposed into intermediate modelling steps, mixing qualitative and quantitative analysis. Section 3 develops an application of the modelling process, within two subsystems of the considered industrial IS considered: the Manufacturing Execution System (MES) and its Real-Time Dispatching (RTD) supporting system. It shows how several meaningful variables for both production and IT managers could be linked to the activity level of physical IT servers, through statistical models. By way of conclusion, section 4 explores and discusses aspects of the on-going research.

2. MODELLING BACKGROUND & PROCESS

2.1 Modelling background

As stated earlier, the goal of this study is to develop an IT capacity planning model for a semiconductor production plant. To build an integrated model, three domains of IT architecture have to be considered: business (industrial), applications and technologies, including hardware (Aerts and al., 2004). Regarding the extent and complexity of those domains, both from IT (Martin and al., 2010) and business (Cooper and al., 1992) levels, a pragmatic and straight to the point approach must be used. Indeed, the historical analytic and simulation computers engineering IT capacity modelling approaches are poorly adapted to the modern industrial IT system, highly distributed, complex and rapidly-changing (Jain, 1991; Gunther, 2007). That is why modelling process should be thought in the context of more recent IT capacity planning studies, focused on trend analysis. These have become increasingly used since about the year 2000 (Gunther, 2007).

To construct such a model, the primary source of information is the so-called “capacity management information system” (ITIL v3 terminology). This information system should centralize all information necessary to manage IT capacity: IT activity (resources consumption, transactions counts, etc.) and, in theory, useful business activity. Setting up this information system is the first step of any ITIL Capacity Management process (Kloosterboer, 2011) and is recommended throughout computer engineering literature (Jain, 1991). The main difficulty is to furnish it with relevant business variables, as IT managers might not be aware of those business activities that impact the IT system. This is typically the case in the research field considered in this paper (STMicroelectronics semiconductor production plant).

Already, semiconductor production systems gather extensive and continuous volume of production activity data. These represent potentially essential sources of information about the business activity. Consequently, it can be argued that this mass of information offers a real and promising opportunity to build predictive causal unified business/IT capacity models, if it becomes possible to select relevant variables and understand their relationships, through a suitable modelling process.

2.2 Modelling process

Confronted to this vast mass of available data and variables, a structured modelling procedure is required, to efficiently converge towards useful and practical models. A bottom-up macroscopic process is suggested. It starts from the predicted IT resource and sets its sight only on the main influencing variables. Two modelling sequences are presented to find out these influencing variables: exploration and explanation. These sequences are divided into several modelling steps, mixing qualitative and quantitative work. Qualitative work is accomplished by involving the experts, through task groups. According to the sequence, quantitative work uses exploratory statistics or forecasting methods.

Sequence 1: Exploration

Step 1.1. Initialization

Data related to a variable, for which a capacity model has to be built, are extracted, eventually corrected, and visualized.

Step 1.2. Depicting the application layer

Applications and sets of software, which are supported by the IT resources considered, are listed with the help of the experts.

Step 1.3. Expressing the variability of the application layer

The assumption here is that the application activity is the principal forecaster for the IT resource activity. As application layers might encompass hundreds of interrelated variables, it is crucial to have a synthetic overview of this activity, to grasp the main sources of variability and summarize relationships between application variables. Multivariate analyses are used for this.

Sequence 2: Explanation

Step 2.1. Finding explanatory business-related variables

The outcome of step 1.3 is the construction of synthetic variables, expressing the application layer variability (usually called “components” or “latent variables”, according to the kind of multivariate analyse used). In this context, such synthetic variables can be considered as underlying unobserved constructs, consequences of the plant’s industrial activities. Thus, these synthetic variables are discussed and interpreted with the experts. When potential explanations are found, measurable explanatory variables (via capacity management information system or mined in the production activity databases) should be put into relation. Such variables should be as much as possible foreseeable as well as understood by production managers.

Step 2.2. Predictive modelling

The identified variables (IT resource and measurable explanatory variables) are quantitatively put into relation, through a statistical model.
Step 2.3. Validation and improvement
The model is presented and confirmed with the experts. Possible improvements are discussed.

3. MODELLING PROCESS: APPLICATIONS
In this section, six steps of the modelling process are applied to two IS subsystems of a semiconductor plant. The first is called MES. It is the backbone of wafers production, involved at each milestone of the production process. The second is the RTD system. It supports MES at some stages of the production process, when they are executed through an automated manufacturing mode. Indeed, two main manufacturing processes are encountered in this production plant: 1) manual, where Work-In-Progress (WIP) dispatching and production lots preparations are done by users and 2) automated, where these actions are accomplished by the IS.

Quantitative analyses were effectuated from autumn 2010 to the summer 2011, using the software package R 2.13.1 (R Development Core Team, 2011).

Sequence 1: Exploration
The objective of this sequence is to express the main sources of variability of the IT applications, with the supposition that they may have an impact the IT resource consumption.

Step 1.1. Initialization
The examined resources are the MES and RTD physical servers. A capacity model has to be built for the total CPU time dedicated to the MES and RTD users. Resources use is expressed in seconds consumed per day. These variables are denoted $\gamma_{MES}^t$ and $\gamma_{RTD}^t$, where $t$ is the temporal scale. For the sake of brevity, the progression over time of these variables is not plotted in this paper.

Step 1.2. Depicting the application layer
This paper focuses on two main applications supported by the servers: 300Works (core MES) for MES and RTD rules manager for RTD. Note that several secondary applications exist on both servers, but depicting 300Works and RTD rules manager activities can also grasp the activities of the satellite applications. 300Works and RTD rules manager activities are tracked in the capacity management information system, by recording their daily volume of transactions calls.

Step 1.3. Expressing the variability of the application layer
To fathom the behaviour of the application layer, the total number of transactions called of both 300Works (445 active during period studied) and RTD rules manager (102 active transactions) is taking into account through multivariate studies.

Firstly, a standard Principal Component Analysis (PCA) is completed (Lebart and al., 2006) for both applications. The analysis is affected by hidden correlations between the synthetic variables (the “components”), due to factor effects (working-days vs. week-ends). Consequently, to get rid of correlations issues, week-end days are removed from the dataset and a new analysis is done. Fig. 1 plots an example of an output of this analysis, for the RTD system: the projections of the original variables onto the two first PCA synthetic variables (factor scores). Most characteristic original transactions are represented, and the IT resource use (CPU.RTD) is added as illustrative variable.

![Fig. 1. Standard PCA: factor scores with some characteristic transactions.](image)

For both applications studies, the four first PCA components predict fairly well the IT resource variables. However, in the MES case, they express less than 50% of the total variability of transactions. This could be explained by the fact that a large number of transactions are used only occasionally. Thus, they are not represented by the main synthetic variables.

To improve the quality of representation of the original variables and the accuracy of the synthetic variables, additional analysis is required. Firstly, a Varimax rotation is applied on the standard PCA axes (ibid.). Secondly, a robust PCA algorithm is applied to the original variables (Hubert and al., 2005), to overcome the influence of potential outliers.

The construction of such synthetic variables, characterized by their representativeness of the application transactions, is the final outcome of sequence 1. The next steps deal with building an explanatory model, to give sense to the statistical observations. For instance, some PCA components reveal an opposition between several application transactions. As it will be described through the application of the sequence 2, these opposed groups of transactions are characteristics of the two concurrent manufacturing processes respectively (table 1).

Sequence 2: Explanation
Step 2.1. Finding explanatory business-related variables
Statistical observations are shared and discussed with experts. First, they explain the functions of the most represented transactions. Based on this, it is possible to describe and understand the functional phenomenon underlying to the synthetic variables.

Then, some explanatory variables, which may explain the activity of the identified functional phenomenon, should be sought. Again, within reason these variables should be understandable and foreseeable. Not surprisingly, as the two
considered subsystems are involved in the wafer production process, they share several explicative functional phenomena. Table 1 summarizes some outputs of the task group with experts: main functional phenomenon and related potential explanatory variables (only synthetic variables used for modelling are presented).

<table>
<thead>
<tr>
<th>Synthetic variable</th>
<th>Underlying functional phenomenon</th>
<th>Explanatory variable proposition</th>
</tr>
</thead>
</table>
| 1 (MES & RTD)      | Opposition between the two main manufacturing processes used within the plant: manual vs. automated. | * Number of production lots prepared through manual mode per day ($V1_t$).  
* Number of production lots prepared through automatic mode per day ($V2_t$). |
| 2 (MES & RTD)      | Overall volume of activity of the plant. | * Number of effective production lots steps performed per day ($V3_t$).  
* Level of WIP (Work-In-Progress) daily snapshot, at a given time ($V4_t$). |

Step 2.2. Predictive modelling

As all quantitative variables are more or less related to the overall activity of the plant, cross-correlations are identified. Some technical operations are needed to limit multicollinerarity, which may bring to an unreliable model. Some outliers, identified as maintenance periods for the IT system, are removed. A first model is computed, using ordinary least squares (ibid.). All explanatory variables are statistically significant (with the exception of the intercept for RTD). Almost 80% of $V1_{mes}$ and more than 85% of $V1_{RTD}$ variability are explained.

Regressions residuals meet the linear model assumptions, with the exception of the independence hypothesis. An analysis of their autocorrelation reveals an AR(1) structure for both models (ibid.). This means ordinary least squares are not adapted to estimate models’ parameters. This glitch is overcome by computing a regression with ARIMA errors (ibid.). Thus, the final forms of the models are:

$$V1_{mes} = \beta_0 + \beta_1(V1_t + V2_t) + \beta_2V3_t + \beta_3V5_t + \beta_4DUM^{WE}_{mes} + \mu_{mes}$$

with $\mu_{mes} = \sigma_{mes}\mu_{mes} + \epsilon_{mes}$

$$V1_{RTD} = \gamma_1(V1_t + V2_t) + \gamma_2V5_t + \gamma_3TIME_t + \gamma_4DUM^{SUN}_{RTD} + \mu_{RTD}$$

with $\mu_{RTD} = \sigma_{RTD}\mu_{RTD} + \epsilon_{RTD}$

where $\beta_i$‘s and $\gamma_i$‘s are the coefficients of the identified explanatory variables for respectively MES ($\beta_0$ is the intercept) and RTD models. $\sigma$ denotes the models’ error terms AR(1) autocorrelation and $\epsilon_i$ the residuals, which are supposed to be independent and identically distributed according to a law $N(0, \sigma^2)$.

Hence, it can be seen through the MES model that the MES cluster time consumed per day ($V1_{mes}$) is estimated through three high-level business variables: the volume of production prepared in manual ($V1_t$) and automated ($V2_t$) modes, and an estimate of the throughput of the plant ($V3_t$). Moreover, a seasonal week-end effect is modelled ($DUM^{WE}_{mes}$) and some unknown phenomena are taken into account, through an additional autoregressive process ($\sigma_{mes}$).

$V1_t$, $V2_t$, and $V5_t$ explanatory variables are shared with the RTD model. On the difference of MES, additional effects due to automated activities deployment are caught by using time as an explanatory variable ($\gamma_1 TIME_t$ instead of $\beta_2 V2_t$). This is necessary for now, to model the ever increasing deployment of automated features in the plant (both in level of activity and new IS features). Lastly, a seasonal Sunday effect is modelled ($DUM^{SUN}_{RTD}$) and some unknown phenomena are taken into account, through an additional autoregressive process ($\sigma_{RTD}$).

Table 2 provides coefficient estimates for models’ parameters.
Table 2. Coefficients estimates for models’ parameters.

<table>
<thead>
<tr>
<th>Variable</th>
<th>MES: $\beta_i$ estimate (standard error)</th>
<th>RTD: $\gamma_i$ estimate (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>161399 (41513)</td>
<td>-</td>
</tr>
<tr>
<td>($V_1t + V_2t$)</td>
<td>58 (3)</td>
<td>122 (1)</td>
</tr>
<tr>
<td>$V_2t$</td>
<td>-31 (4)</td>
<td>-</td>
</tr>
<tr>
<td>$V_5t$</td>
<td>13 (2)</td>
<td>19 (5)</td>
</tr>
<tr>
<td>$TIME_t$</td>
<td>-</td>
<td>2401 (108)</td>
</tr>
<tr>
<td>$DUM_{t+RE}$</td>
<td>-31529 (2910)</td>
<td>-</td>
</tr>
<tr>
<td>$DUM_{t+SUN}$</td>
<td>-</td>
<td>-27949 (7219)</td>
</tr>
<tr>
<td>AR(1) process</td>
<td>0.5 (0.06)</td>
<td>0.6 (0.05)</td>
</tr>
</tbody>
</table>

As represented in fig. 2 and 3, only a few points are outside the 5% confidence intervals: the models represent pretty well the subsystems’ real behaviour.

![MES real data versus model fitted values](image1)

Fig. 2. MES model: real data and fitted values

![RTD real data versus model fitted values](image2)

Fig. 3. RTD model: real data and fitted values

Furthermore, predictive features of these models are positively verified with time-series cross-validation algorithms (Hyndman, 2011). For both models, the forecast mean absolute error is always lower to 5% of the average of the observed CPU consumption on the period.

Step 2.2 deals with building two models, which seems to be valid. After having presented the final forms of the statistical models, it has been demonstrated that fitted values are close to real data. Finally, a positive predictive test has been performed on the whole dataset. This indicates that, at this stage of the modelling process, results are valid and encouraging. This demonstrates the feasibility of making reliable predictions of IT variables’ behaviour, on the basis of explanatory variables to represent the business activity. Applying a systematic and rigorous modelling process makes this achievement possible.

Step 2.3. Validation and improvement

A final consultation meeting is setup with the experts. They agree upon the relevancy of MES and RTD models, and improvements are discussed. This won’t be detailed in this paper. However, by way of example, the case of automated activities impacts will be briefly regarded.

Through the MES model, it can be observed that an increase of the volume of automated activities ($V_2t$) will have, all other being equals, less impact on the MES server than an increase of the volume of manual activities ($V_1t$). This is shown through the negative $\beta_2$ coefficient. CPU consumption data confirm this statement: despite the fact that they are supporting similar explanatory variables, it can be observed that the MES IT resource consumption remains quite stationary. This is in contrast to the RTD consumption, which is increasing steadily on the period studied. This makes sense, as some manufacturing activities are transferred from the MES server to the RTD server when automated manufacturing replaces manual processing. As automation is used more and more often, this is the first primary cause of increased consumption for the RTD server. Of secondary importance is the constant deployment of new automated IS features, which generate increasingly additional loads on the RTD server. Unfortunately, the impact of the automated activities on the RTD server cannot be evaluated with precision, due to the deployment of these new IS features on the period studied. However, when and if the system stabilizes, an update of the RTD model can be possible to develop, in order to ingrate $V_2t$ as an explanatory variable, instead of $TIME_t$.

This ends the sequence 2 and the modelling process.

4. CONCLUSION & PERSPECTIVES

This paper has shown how to link the utilizations of IT variables to several manufacturing variables, by applying a systematic and rigorous modelling process. An actual application on two STMicroelectronics IT subsystems demonstrated the generic nature of the modelling process. Moreover, differences and similitudes were detected between the subsystems. This leads to an interesting future research perspective: the identification of common behaviour throughout the IS.

Before this, section 3, step 2.3 concluding remarks brought some insights that are going to be considered in the near future, in effect to ground the modelling process within STMicroelectronics’ industrial context. First, it stated that the MES model may have to be revaluated later. That is potentially true for all models built with this modelling
process. Indeed, the assumption behind all forecasting models is that statistical inferences about the present or future are made using historical data. It means that the model of a system will only be true if the system under consideration remains structurally stable. As the semiconductor and other industrial environments are fast-changing (on both industrial and IS/IT levels), a one-shot static modelling process is not sufficient. The relevancy of the model over time should be checked, in order to be able to update it in case of structural change. We are currently working on improving the modelling process with an online model monitoring (Zeileis and al., 2005).

Then, the question of using the statistical models within a complete decision-making tool needs still to be studied in greater detail. What matters here is the capability to integrate statistical models within a full industrial scope, as mentioned in the introduction. Truly, the final objective of this research is to design a thorough explanatory system, linking industrial variables, organized according to industrial scenarios, to several potential supporting IT resources. The results presented in this paper are only a first step toward such a more systemic view of the industrial processes. Firstly, business operations have to be considered with deeper precision, to be fit into the scheme of an industrial use of the Information System, i.e. according to relevant industrial scenarios. Explanatory variables of MES and RTD models might not be the only business key indicators used to evaluate business scenarios. Consequently, business variables that are worthwhile for valuable industrial scenarios must be understood and integrated in a comprehensive capacity model. Secondly, MES and RTD models showed how common explanatory variables may impact upon different IT resources. Moreover, it was discovered that an industrial variable may have, all other things being equals, opposite effects on different IT resources. In the case studies, it has been concluded that an increase of automated activities ($V_2$) will diminish the load for the MES server, but increase the workload of the RTD server. This means that the capacity failure risk is moved from one resource to another. Thus, a comprehensive capacity model should cover the evaluation of such balancing between IT resources. We think that such a general quantitative predictive model, accurate enough on both business and IT layers for relevant industrial scenarios, could be achieved by applying our modelling process on several points of the industrial IS. Further research will confirm or refute this proposition.

ACKNOWLEDGEMENTS

The authors would like to thank STMicroelectronics for supporting this research project and Marie-Agnès Girard for her inspiring insights and Chris Yukna for his help in proofreading.

REFERENCES


