A Procedure and Program to Calculate Shuttle Mask Advantage

A.Balasinski[1], J. Cetin[1], A.Kahng[2], X. Xu[2]

[1)]Cypress Semiconductor, 12230 World Trade Dr., San Diego, CA 92128, USA
[2)]University of California, San Diego

ABSTRACT

A well-known recipe for reducing mask cost component in product development is to place non-redundant elements of layout databases related to multiple products on one reticle plate [1,2]. Such reticles are known as multi-product, multi-layer, or, in general, multi-IP masks. The composition of the mask set should minimize not only the layout placement cost, but also the cost of the manufacturing process, design flow setup, and product design and introduction to market. An important factor is the quality check which should be expeditious and enable thorough visual verification to avoid costly modifications once the data is transferred to the mask shop. In this work, in order to enable the layer placement and quality check procedure, we proposed an algorithm where mask layers are first lined up according to the price and field tone [3]. Then, depending on the product die size, expected fab throughput, and scribeline requirements, the subsequent product layers are placed on the masks with different grades. The actual reduction of this concept to practice allowed us to understand the tradeoffs between the automation of layer placement and setup related constraints. For example, the limited options of the numbers of layer per plate dictated by the die size and other design feedback, made us consider layer pairing based not only on the final price of the mask set, but also on the cost of mask design and fab-friendliness. We showed that it may be advantageous to introduce manual layer pairing to ensure that, e.g., all interconnect layers would be placed on the same plate, allowing for easy and simultaneous design fixes. Another enhancement was to allow some flexibility in mixing and matching of the layers such that non-critical ones requiring low mask grade would be placed in a less restrictive way, to reduce the count of orphan layers. In summary, we created a program to automatically propose and visualize shuttle mask architecture for design verification, with enhancements to due to the actual application of the code.

Keywords: Masks, Reticles, Multi-Layer, Multi-Product, Mask Cost, Shuttle, IP

1. INTRODUCTION

New product development includes many risks, related to technology, design, and market conditions. Semiconductor companies are reducing the cost of this risk, by economizing on the initial mask set needed to prove the design. This methodology pertains mostly to low-volume testchips consisting of subsets of design databases (intellectual property, IP). The savings on the IP placement on the mask set can be substantial, e.g., corresponding to cutting the $1M price tag per product for 90 nm technology by several times [4]. A standard implementation to place non-redundant components of the IP on one reticle plate [1, 2] is to create reticles known as multi-product, multi-layer, or multi-IP masks [3]. While conceptually, it is a simple exercise, in practice, one has to consider a variety of options using careful calculations. These calculations should be expeditious and enable immediate quality check, preferably by visual inspection. Verification of multi-IP mask architecture is more critical compared to the standard, single product mask sets because, once the data is transferred to the mask shop and the first reticle is shipped, it is very costly to change it, significantly more than in the conventional scheme of one product or layer per plate. In this work, in addition to the fully algorithmic approach, we show how the visual scrutiny provided us with the opportunity of making some common-sense choices which we further developed into mask placement rules. This way, we not only ensured that the layer placement on the mask is correct, but we also enhanced its final architecture.

2. RETICLE COMPONENT IN PROJECT COST

The basic rules of layer placement on the reticle set correlate layer types with mask grades. This means that the reticle quality (and therefore, the price) as well as the field tone (clear, opaque, PSM) have to correspond to the drawn layer characteristics (normal or reverse drawing, e.g., lines vs. contact holes), overlay requirements, the existence of sub-resolution assist features (SRAF), etc.. In every process, there are a number of design layers with identical or similar properties which would make it possible to combine their databases on one reticle and reduce the overall project cost. Indeed, beginning from the 130 nm technology node, the placement of one layer (L) of one product (P) on one plate, (single IP, 1L-1P masks) was found not to be sufficiently cost efficient. The approach proposed instead, the multi-IP masks, e.g., NL-1P (N = number of layers) or 1L-MP (M = number of products) aimed at the up-front reduction of the project development cost. However, in reality, the changes in mask architecture had much more profound consequences, most of which were driving up other components of project cost. Accordingly, it became clear that to modify mask structure, one needs to consider tradeoffs with setup complexity and the throughput of design, product, or manufacturing lines. Eq. 1 shows the key components of the total cost of product development process *TOTAL(IPD)*:

$$TOTAL(IPD) = RET(IPD) + MFG(IPD) + SET(IPD) + DES(IPD) \quad (1)$$

where

- *RET* is the up-front cost of IP placement on reticles, some of which is repetitive with the subsequent tapeouts,
- *MFG* is the distributed cost of wafer manufacturing depending on the mask scheme in the fab,
- *SET* is the up-front cost of design and manufacturing tool setup, which can also repeat with tapeouts,
- *DES* is the distributed cost related to the design/product line execution, including the cost of time to market and design complexity,
- *IPD* is the relative density of design IP on the masks taping out, calculated from the number of layers or products per plate, where *IPD*=1 corresponds to the single IP mask.

The two key cost categories, the up-front and the distributed cost have the following characteristics:

- Up-front cost – single time or repetitive, incurred once at decision points, e.g., at tapeouts with a given mask architecture. While this cost can be amortized over time, it can not be changed.
- Distributed (continuous) cost which is incurred gradually over time. This cost can be modified at any point of time, going forward.

Fig.1 compares the time-dependent cost events for the single-IP and multi-IP masks.

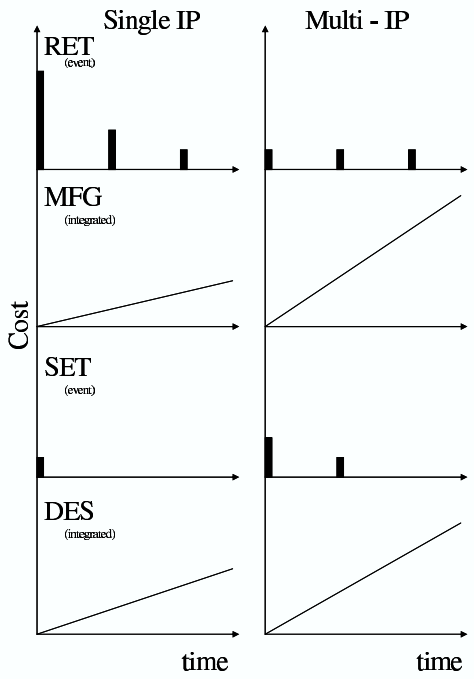

Fig.1. Conceptual time dependent distribution of cost events (step vs. ramp) for new project development using single- and multi-IP masks. The vertical bars correspond to the up-front cost incurred at tapeout, the ramp symbolizes the distributed cost accrued while running wafers in the fab or waiting on silicon data.

Correlation between the *IPD* and the number of

layers (N) or products (M) per plate is given by:

$$IPD = \frac{Total_number_of_masks_in_a_Single_IP_architecture}{Actual_number_of_masks_taped_out_with_the_same_IP},$$

which, for the two basic multi-IP schemes, NL-1P and 1L-MP, can be reduced to:

$IPD = N$, for NL-1P masks, assuming no orphan layers

$IPD = M$, for 1L-MP masks, assuming same flow for all products.

Cost minimization would therefore require multi-parameter optimization, such that first derivative of Eq. (1) by the *IPD* equals zero:

$$\frac{d(TOTAL)}{dIPD} = \frac{dRET(IPD)}{dIPD} + \frac{dMFG(IPD)}{dIPD} + \frac{dSET(IPD)}{dIPD} + \frac{dDES(IPD)}{dIPD} = 0 \quad (2)$$

One should note that, to find the minimum, the value of Eq.2 for $IPD<IPD_0$ should be < 0 and for $IPD>IPD_0$ should be > 0, where IPD_0 is the optimal *IPD* value. Now that the reason behind the multi-IP masks is to reduce the cost of mask information placement as the most important cost component in Eq.(2), by increasing the IPD, it results that:

$$\frac{dRET(IPD)}{dIPD} < 0 \quad (3)$$

One should note that in this simple model, the cost of IP placement on reticles is equal to:

$$RET = \frac{Cost_of_reticle}{IPD}$$

where *Cost_of_reticle = const (IPD),* i.e., does not depend on the information content.

Formula (3) could be brought to the maximal negative values by maximizing the information content of the reticles, i.e., increasing the number of layers or products per plate. However, for $IPD>IPD_0$, the other components of Eq. (2) would start increasing, due to the increased mask complexity, such that:

$$\frac{dMFG(IPD)}{dIPD} >> 0 , \quad \text{or} \quad \frac{dSET(IPD)}{dIPD} >> 0 , \quad \text{or} \quad \frac{dDES(IPD)}{dIPD} >> 0$$

By inserting example numbers into Eq.2 and redrawing Fig.1 accordingly, one can visualize the time dependent cumulative, single- and multi-IP project cost. Fig.2 shows how this cost is sensitive to the additional *SET* and *DES* components which accumulate over time to eventually become comparable to the reduced *RET* cost. This illustrates the need to carefully optimize the setup and design procedures. The 50% advantage in the initial project cost for the multi-IP scenario gradually shrinks and disappears after about 30 weeks, due to the significant cost overhead on the part of the manufacturability, setup, and design, for the multi-IP scenario. This is due to the fact that, even though conceptually simple, multi-IP masks require significant effort in capturing, processing, and storing information related to mask architecture, special setups from design and fab and custom quality procedures. For example, under the 1P-1L scheme, where masks were taped out for one product only, there was no need for special placement algorithms, stepper jobs, revision control, database archiving, etc. With multi-IP masks, one can not easily verify mask names and numbers, and the revisions need to entail both the individual piece of database and the entire reticle set, possibly leading to confusion. These risks are especially important when multiple design groups are involved in keeping track of their databases, possibly to the different sets of standards.

In summary, the key issues increasing the cost of multi-IP mask approach are related to the handling of mask layer placement, complexity reduction, design- and fab-friendliness, and elimination of errors. These issues can be addressed by automated, manual, or mixed approach, as discussed below.

3. MULTILAYER MASK ALGORITHM

The commonly preferred approach of complex data handling is to automate it. Accordingly, for the NL-1P multi-IP mask architecture, we developed an algorithm where the design IP layers were first lined up according to the price and field tone [5]. Then, depending on the product die size, fab throughput, and scribeline architecture, these layers were placed on the masks with different grades. Fig.3 shows the flow chart of the procedure for a single product. The first part

of the procedure, the decision on the NL-1P vs. 1L-1P mask set depends on the risk factors for technology, design, and product marketability which are the parts of *SET* and *DES* cost components. The savings are also impacted by the reduced fab throughput (*MFG*) and are usually about 30% lower than the entitlement related to the *IPD*, due to the matching restrictions and the resulting orphan layers.

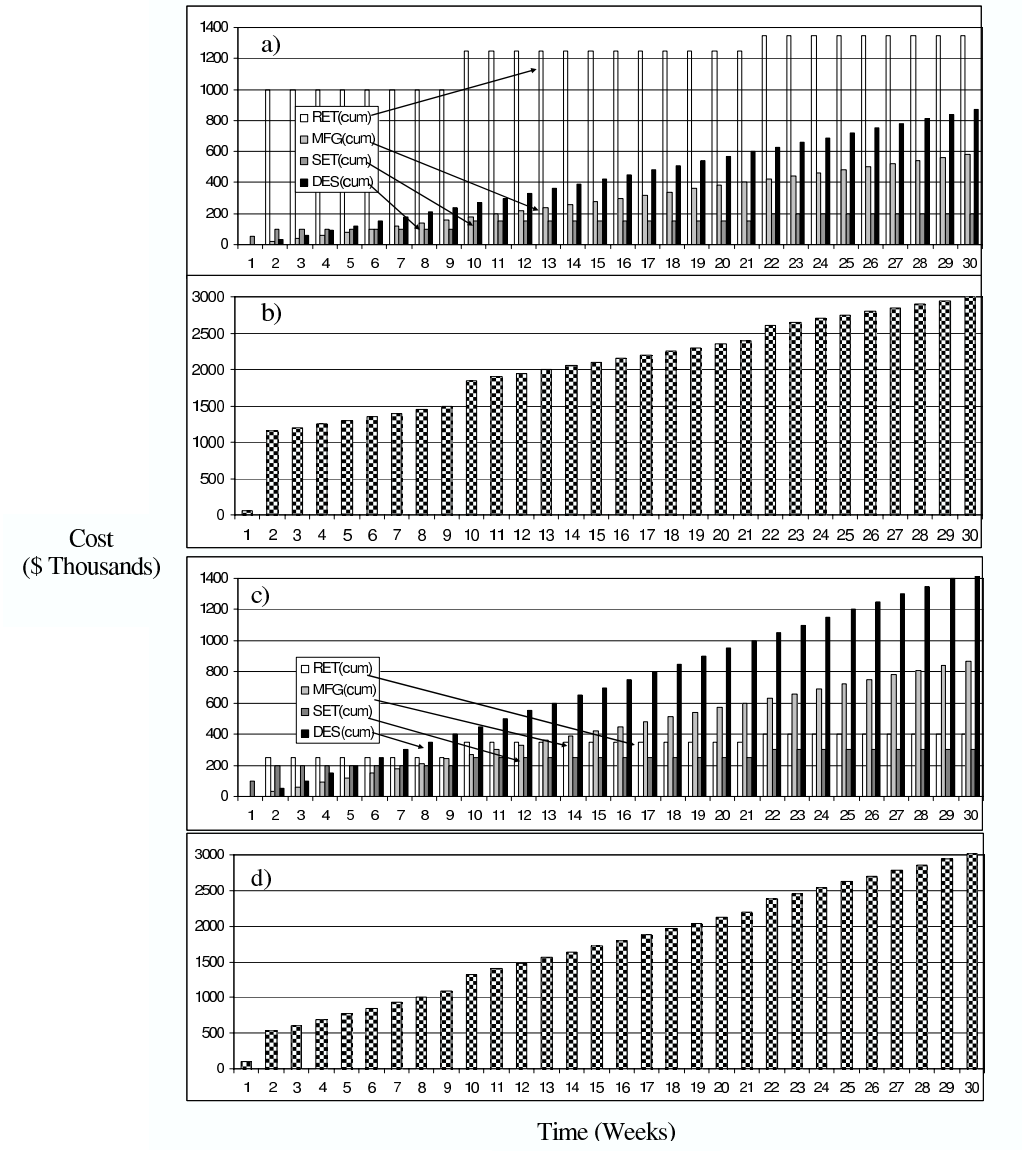

Fig.2. Example of cumulative project cost as a function of product development time, for 1L-1P scheme: a – partitioned into four components: *RET, MFG, SET, DES*, b – combined, and for NL-1P or 1L-MP schemes: c – partitioned into the four components, d – combined.

The second part of the algorithm shown in Fig.3 describes the placement of individual design layers over the reticle field. As indicated in Fig.1, one can expect that mask retapeouts are the intrinsic part of product development, e.g., to improve product performance. While mask cost for these retapeouts could be reduced e.g., according to the NL-1P scheme, this could come at the expense of the setup cost. Tapeout related activities can not be fully automated, or the automation would require complex, expensive procedures. Therefore, significant portion of design verification has to be done manually, adding to the *SET* cost.

Fig. 4 shows the user interface (GUI) for the multi-layer mask planner developed at UCSD based on the algorithm from Fig.3. It first defines mask rules related to the field sizes and types of multi-IP masks. Next step is to input layer information such as mask grades, field tones, SRAF options, etc. The program then automatically generates graphical representation of the mask along with its key parameters and die names or numbers, as shown in Fig. 5 (left). Finally, it shows how the mask is stepped over the wafer area to provide the number of exposures for fab approval (Fig.5, bottom).

It is important to realize technical limitations of such generic multi-IP mask approach as it may happen that that the cost of the setup could become too high as a result of automation rules not verified against fab or design guidelines. For example, manufacturing requirements are such that the valid number of layers per plate is limited to a few (2-4). Higher number would reduce the field size and require excessive number of exposures per wafer. Therefore, one can define technology layer databases with fixed layer sets based on a single run of the algorithm and reuse them only for projects that can benefit from the multi-IP masks. For a small number of layers per plate dictated by the die size and design feedback it may be easy enough to manually identify pairing options and further simplify the placement procedure.

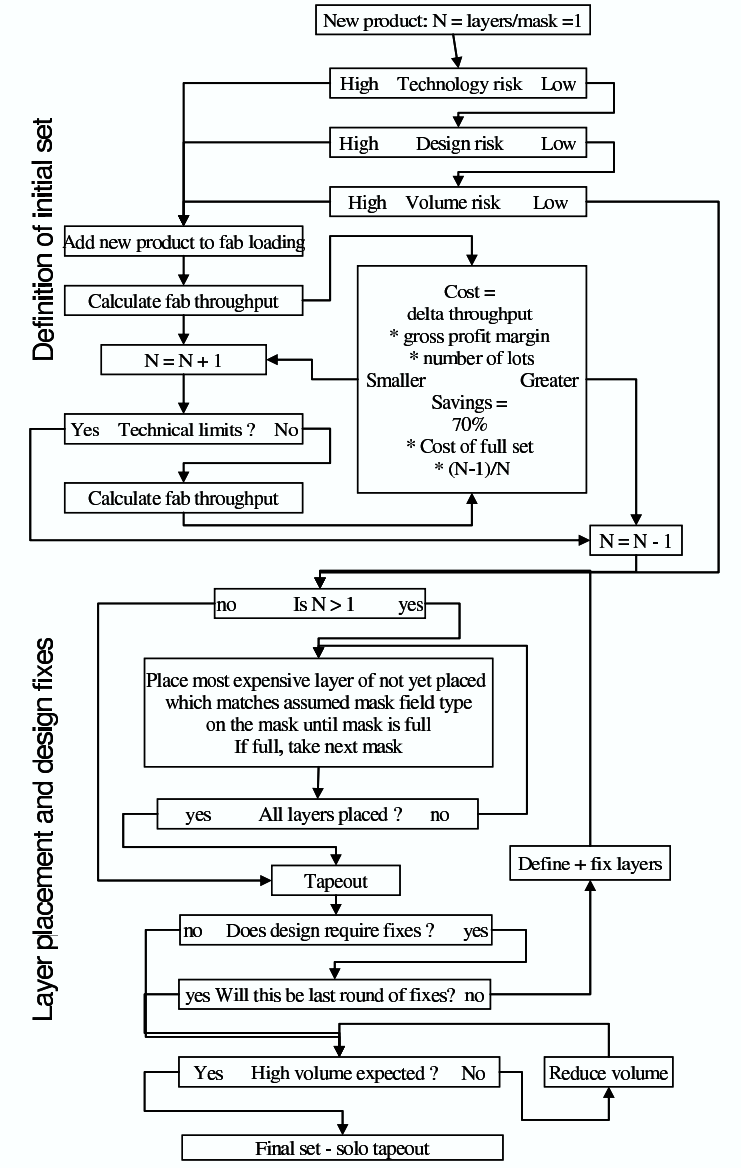

Fig.3. Generic flow chart of multi-IP mask building procedure for a single product.

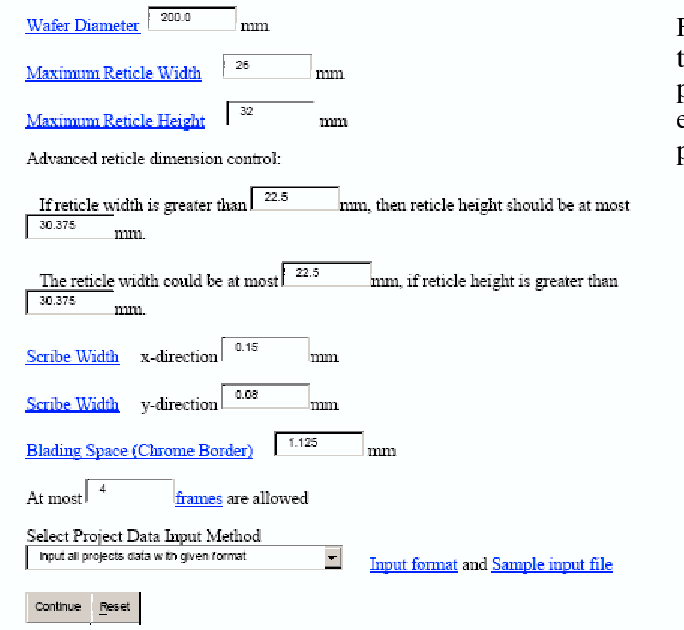

Fig.4. Input GUI template of the multi-IP program, allowing to enter the basic mask parameters.

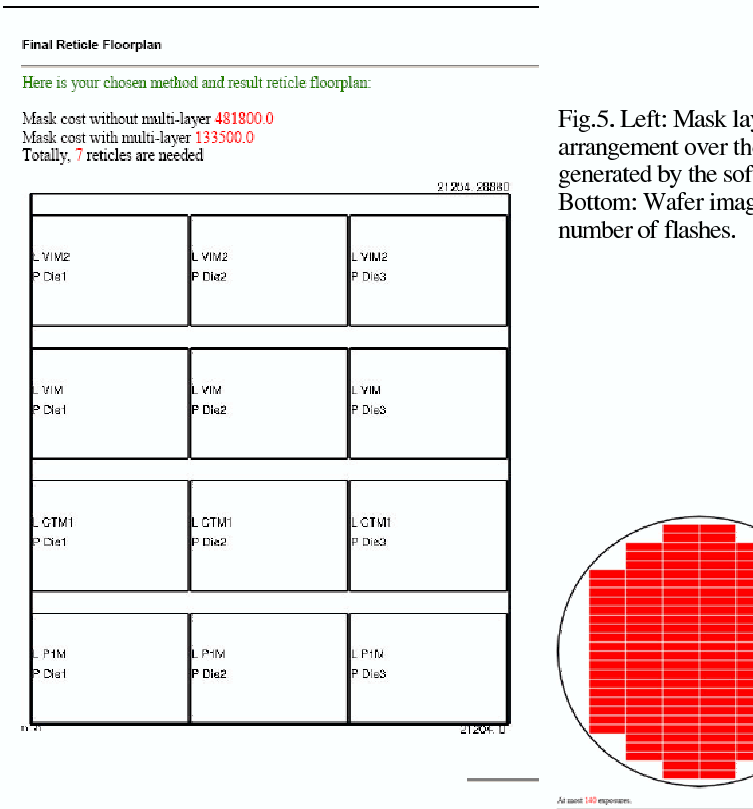

Fig.5. Left: Mask layer arrangement over the reticle field generated by the software. Bottom: Wafer image showing the number of flashes.

4. GENERIC MULTI-IP MASK PLANNING

The other basic multi-IP mask model is the 1L-MP architecture. Its die placement methodology assumes that all products taping out would have identical number of design layers [1]. This is true if these products share the same manufacturing flow. However, product lines may wish to add unrelated databases onto the mask set, to maximize the savings. It appears that a special methodology and infrastructure of database tracking needs to be put in place to ensure error free layer placement.

While the layer placement procedure discussed in previous Section can be used for the majority of the layers, it may not be sufficient from the *SET* and *DES* cost reduction standpoint. One may need to propose enhancements, limitations and additional rules:

- layers with similar functions should be placed on one reticle, to enable single mask retapeout for design fixes,
- non-critical layers can be placed with the critical ones even if their characteristics, e.g., field tone, are different, as per the guidelines provided by the maskshop,
- layers with similar pattern densites should be placed on the same reticle,
- individual layers from different products need large footprint so as not to reduce fab throughput.

These rules may not lend themselves to easy automation. While the basic layer combinations call for price and quality-dependent layer pairing, this may not always be the best solution for product development. Also, manual pairing can help reduce the count of orphan layers. Following this, we enhanced the program to allow for manual pairing as the first option.

As an extreme case of mask complexity, product lines may choose to process all the design information readily available at prescribed times through the mask shop. Here, the design IP layers from different products would need to be lined up to be transferred to the reticles, as shown in Fig.6. The placement methodology would be similar to placing passengers on a plane or containers inside a ship, on first-come, first serve basis, with significant randomness allowed.

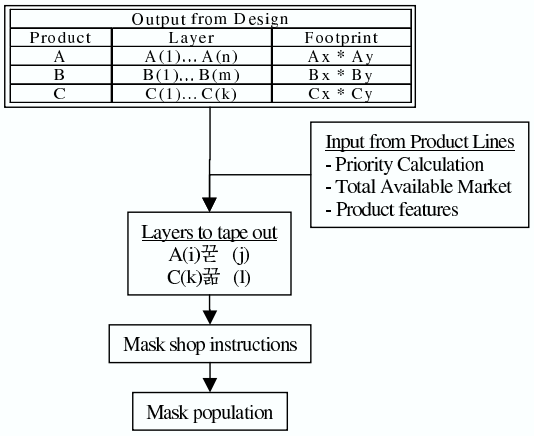

Fig.6. Population process of a multi-IP mask consisting of multiple layers of comparable grades.

The priority and placement may depend on the type of the project. A dynamic placement algorithm needs to act on the mask IP "passengers" as if they were material objects, with different sizes and demands, and load them up on the masks as they come. Standard devices for volume production would follow the customary 1L-1P architecture, whereas testchips for new technology which tend to be custom based and use much more expensive mask set, would vary case by case. Development of layer placement software should satisfy their separate needs.

Generating and reviewing mask shop instructions based on the placement algorithms would be a process of increasing complexity. A question would arise, how many items should be automatically populated and how many of them should be individually dialed in. To answer this question, one may refer back to Eq.2 to see what is the impact of the number of masks on the *SET* of *DES* complexity and time to solution.

5. CONCLUSIONS

Shuttle automation is making significant progress. Tools are becoming available to calculate mask cost reduction and optimize the placement of multiple products on the reticle plate for easy automation, based on their footprint and scribing/sawing requirements. However, even with this progress, the issues are not always becoming simpler. This is because one has to deal now with much more complex data structure as compared to the single product where it was possible to send the data to the mask shop in a small number of packets, corresponding to the die database, frame database, and revision number, along with simple instructions. As a result, the maskshop was able to easily follow these instructions to create the mask architecture. Nowadays, proper placement of layers on the mask needs to be provided by a mixture of manual "common sense" approach and automated rules to mix and match physically uncorrelated layers, minimize the number of orphans and make the set manufacturing-friendly at the same time. We found that algorithmic solutions aimed only at reduction of the mask cost may not always provide optimal mask composition. This is because there are too many custom elements which do not lend themselves to the easy translation into mask design rules.

We showed that overall cost optimization needs to be based on multi-parameter, integrated cost equation. This equation takes into account not only the up-front cost of the mask and the distributed cost of the process but also the cost of setup, product line, risk of error and lost opportunity. It is by combining all these factors that one can propose the best approach which should sometimes be algorithmic and sometimes manual and visual. In any case, this work shows that there is significant cost of the complexity of the solution over the product lifetime, which needs to be included when developing mask architecture.

6. ACKNOWLEDGMENT

We thank Wendy Sachs-Baker for help in verifying the different options of the program.

REFERENCES

1. A.B.Kahng, I.Mandoiu, X.Xu, and A.Zelikovsky, Yield-Driven Multi-Project Reticle Design and Wafer Dicing, Proc. Int'l Symp. on Physical Design, pp. 70-77, April 2004.
2. M.C.Wu and R.B.Lin, Multiple Project Wafers for Medium-Volume IC Production, Proc. IEEE International Symposium on Circuits and Systems, ISCAS 2005, vol.5, pp.4725 – 4728, May 2005.
3. A.Balasinski, Multi-layer and Multi-product Masks: Cost Reduction Methodology, Proc. Photomask Japan, April 13-15, 2005, pp. 91-92
4. www.eetimes.com/news/semi/showArticle.jhtml?articleID=160401464
5. A.Balasinski, S.Larky: patent pending.