Sensitivity Analysis of Multi-project Wafers Production Cost

Ming-Hsine Kuo, Meng-Chiou Wu, and Rung-Bin Lin
Computer Science and Engineering
Yuan Ze University
Chung-Li, 320 Taiwan
s937433@mail.yzu.edu.tw, mcwu@vlsi.cse.yzu.edu.tw, csrlin@cs.yzu.edu.tw

ABSTRACT

Multiple-project wafer (MPW) is now indispensable to mask cost reduction for low-volume integrated circuit production. Due to the high uncertainty of project design progress, a project may be abruptly withdrawn from or added to an MPW run. In this paper, we perform a cost sensitivity analysis of withdrawing a project from or adding a project to an MPW run. The analysis is especially useful for a foundry or a design service company to decide whether to accommodate one more project in an MPW run. The experimental results show that removing/adding a project with large chip area would increase/decrease the cost shared by each individual project for low-volume production. However, as production volume increases, no single rule can be drawn to characterize the change of the shared cost of each project. This further gives evidence to the need of having a systematic approach like the one proposed in this paper to evaluating such a change in an MPW run.

1: INTRODUCTIONS

As semiconductor process technology advances into the deep sub-micron era, mask cost is skyrocketing [1]. At the 90 nm node, the mask cost for an ASIC is about 1 million dollars, compared to $300,000 for a 180 nm node [2]. This is due to the wide use of resolution enhancement techniques such as optical proximity correction and phase shifting mask for sub-wavelength lithography. To curtail the enormous increase in mask cost for chip fabrication, multi-project wafer (MPW) or shuttle run is used to reduce the cost for low-volume integrated circuit (IC) production [3]. By doing so, the projects participating in an MPW run can share the mask cost and each of the projects will thus pay a lower mask expense.

Despite of being as an important vehicle for low-volume IC fabrication, an MPW run is subject to the situation where a project is withdrawn abruptly because the project can not be finished timely to take the scheduled shuttle. It is also subject to the situation where a project is added abruptly because the project is completed much earlier. In this paper, we are interested in finding how sensitive the total fabrication cost of an MPW run and the cost assumed by each individual project are to the change of the projects participating in the MPW run. The sensitive analysis is especially useful for a foundry or a design service company to decide whether to accommodate one more project in an MPW run. To perform a robust analysis we carry out a design space exploration to find out a minimum cost solution prior to and subsequent to adding/removing a project from/to an MPW run, respectively. Our experiments show that removing a project with large chip area would increase the cost shared by each individual project for low-volume production. However, as production volume increases, no single rule can be drawn to characterize the change of the shared cost of each project. This is due to none of the production cost components - mask cost, exposure cost, and field-size independent wafer cost, dominates the total fabrication cost for higher volume production. Therefore, we need a systematic approach like the one proposed in this paper to performing a sensitivity analysis. Herein we will use chip and project interchangeably except that it is indicated otherwise.

The rest of this paper is organized as follows. Section 2 presents a problem description. Section 3 describes the cost models used in our study. Section 4 describes our approach to the cost sensitive analysis of an MPW run. Section 5 gives some experiments. Section 6 draws a conclusion.

2: PROBLEM DESCRIPTION

We are to study how sensitive the total fabrication cost of a shuttle run and the cost assumed by each project are to the change of the projects participating in the shuttle run. To perform such a study, we need first a cost model for computing mask cost and wafer production cost. Mask cost can be easily estimated if we know the reticle (mask) area used for the underlying shuttle run. On the contrary, it is difficult to estimate the wafer production cost because it is difficult to know in advance the number of wafers needed to be fabricated. The reason for this is as follows. Since the chips participating in a shuttle run normally have different widths or heights, we face a problem of arranging them in a reticle so that the number of required wafers can be minimized. The problem of arranging the chips in a reticle is called reticle floorplanning or shuttle mask floorplanning [4-6]. The result of reticle floorplanning is called a reticle floorplan. Fig. 1 shows a reticle floorpan used for fabricating an MPW [6]. There are 10 projects (chips) participating in the MPW run.

Given a reticle floorplan, we need determine the wafer dicing plan for each of the wafers to know the number of wafers needed to be fabricated. A wafer dicing plan consists of a set of reticle dicing plans, each of which in turn consists of a set of horizontal and vertical dicing lines being applied to a reticle to obtain some bare dice. For example, the lines v_1, v_2, v_3, h_1, h_2, and h_3 form a reticle dicing plan which would produce good bare dice for chips 5 and 8. Finding wafer dicing plans that turn out a minimum use of wafers is a difficult problem due to side-to-side dicing constraint. This constraint requires that a dicing line start from one side of a wafer and stop at the other side of the wafer. This creates dicing conflicts among chips, i.e., dicing one chip out may destroy several other chips.

Even if we could find out a reticle floorplan that results in a minimum use of wafers, this does not mean we would attain a minimum cost solution. Mask tooling cost, wafer exposure cost, and wafer processing cost other than exposure cost all together determine the total production cost of an MPW run. Thus, a design space exploration need be done to find a minimum cost solution rather than a solution with a minimum use of wafers.

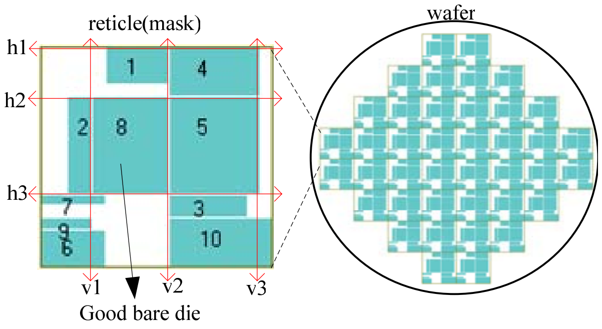


Fig. 1. A reticle floorplan and an MPW.

3: MPW PRODUCTION COST

The total cost of an MPW run includes mask tooling cost and wafer fabrication cost. Wafer fabrication cost is further divided into wafer exposure cost and field-size independent wafer cost [7,8]. Given a reticle floorplan with area $\Re$, the total MPW fabrication cost is

$$T_{mpw}(\Re) = C_m(\Re) + Q(\Re)C_e(\Re) + Q(\Re)C_w , \qquad (1)$$

where $C_m(\Re)$, $C_e(\Re)$ and C_w are the mask cost, exposure cost per wafer, and field-size independent wafer cost per wafer, respectively. $Q(\Re)$ is the number of wafers fabricated. Note that exposure cost per wafer is a function of wafer filed size, but the cost per exposure is a constant.

3.1: MASK TOOLING COST

Mask cost is mainly incurred by data preparation, mask write, mask inspection, mask repair, etc [9,10]. The materials used for mask tooling also contribute a considerable portion of mask cost, especially for advanced technology nodes. Mask yield highly depends on the number of (very) critical layers used in a chip and the total area of the chips in a reticle. It should be a non-linear function of total chip area as shown in Table 1. The data in Table 1 are calculated for a 90nm technology node assuming that a chip has 8 very critical layers, 8 critical layers, and 12 non-critical layers [9]. They are originally estimated for a reticle containing a number of dice of the same design with 8*8 mm^2 wafer field size (defined later). Note that the mask cost for an MPW run should be somewhat larger than that given in table 1 due to more data preparation time.

Table 1. Mask cost for different wafer field sizes.

Wafer field size	25*25 625 mm^2	16*24 384 mm^2	16*16 256 mm^2	8*16 128 mm^2	8*8 64 mm^2
Mask cost	1,240,000	728,000	532,000	352,000	296,000

3.2: WAFER COST

The main contributors to wafer cost include exposure, hot process, etch, sputter, polish, etc [10]. Among them, exposure cost highly depends on the type of layers employed in a chip. The cost per exposure for a very critical layer can be five times that for a non-critical layer [9]. Exposure cost per wafer also depends on the number of reticles printed on a wafer, i.e., depends on the wafer field size. This part of cost is also called *field-size dependent wafer cost*. The part of wafer cost other than exposure cost is called *field-size independent wafer cost* which is independent of field size, i.e., the number of reticles printed on a wafer. *Field size* is the size of an area on a wafer exposed to the light during per reticle exposure. It is normally 1/4 or 1/5 times the corresponding reticle, depending on the projection lens. In this paper, what we mean reticle size is its corresponding field size except that it is indicated otherwise. Once we know the number of wafers fabricated, we can calculate total wafer fabrication cost using the last two terms in (1).

3.3: COST SHARING MODELS

Cost sharing model is used to compute the production cost assumed by each individual project. It must render a fair share of cost for each of the projects in an MPW run. Our cost sharing model is based on the model proposed in [8]. The model in [8] is the first reasonable cost sharing model if the production volumes of the projects are all similar. The total cost for a project p going with an MPW run whose reticle has a wafer field size $\Re$ is

$$\begin{aligned} C_{mpw}(p) = {} & C_m(\Re) A_p \Big/ \sum_{i=1..N} A_i + \\ & C_e(\Re) Q(\Re) V_p \Big/ \sum_{i=1..N} V_i + \qquad (2) \\ & Q(\Re) C_w A_p V_p \Big/ \sum_{i=1..N} A_i V_i \end{aligned}$$

where A_p, V_p and N are the area of chip *p*, the production volume of chip *p*, and the number of projects in an MPW run. The first term in (2) is the share of mask cost. The second term is the share of exposure cost. The third term is the share of field-size independent wafer cost. When all chips have the same production volumes, the share of field-size independent wafer cost is proportional to chip area and exposure cost is evenly shared. As production volume varies, this cost model tends to favor the projects with larger production volumes. The reason is that (2) gives the projects with smaller production volumes a share of the cost of fabricating the wafers solely used for the projects with larger production volumes. This results in overcharging the project with smaller production volume.

We here propose a fairer shared cost calculation approach. The main idea behind this approach is as follows. We create several volume floors based on the production volumes required by the projects in ascending order. We then use (2) to calculate a shared cost for a project based on the volume requirement between two adjacent volume floors. Then, the total shared cost of a project is the sum of all the shared costs incurred from volume floor *i* to floor *i*+1 for all *i*. For the example shown in Fig. 2, we assumed that there are four chips {*B*, *C*, *D*, and *E*} in an MPW run and their required volumes are 20000, 100, 200, and 50000 dices, respectively. We create four volume floors based on the given production volumes. The first volume floor needs $Q_1(\Re)$ wafers to satisfy a volume requirement of 100 dice for projects *C*, *D*, *B*, and *E*, respectively. The second volume floor needs $Q_2(\Re)$ wafers to satisfy a volume requirement of 100 dice for project *C* and 200 dice for projects *D*, *B*, and *E*, respectively. Thus, it needs $Q_2(\Re)-Q_1(\Re)$ wafers to produce extra 100 dice for projects *D*, *B*, and *E*, respectively. A similar interpretation can be applied to the other pairs of adjacent volume floors. Thus, using (2) to compute the shared cost from volume floor *i* to floor *i*+1 we need only consider those projects that have volume requirements exceeding the volume at volume floor *i*. For example, project *C* basically should not assume the cost of fabricating extra $Q_4(\Re)-Q_1(\Re)$ wafers. Similarly, project *D* will not share the cost of fabricating extra $Q_4(\Re)-Q_2(\Re)$ wafers. As one can see, this approach will compute a fair shared cost for each of the projects in an MPW run.

4: SENSITIVITY ANALYSIS

It is clear that removing a project would certainly not increase the total MPW production cost. However, it is totally unclear how the cost shared by each project would change due to removing a project from an MPW run. Assumed that all projects have the same production volume and the same chip area, we still can not say anything about the shared cost of a project even if the total mask cost, exposure cost, and field-size independent wafer cost are all reduced after a chip is removed. The reason is that removing a project under such an assumption increases the portion of the mask cost, exposure cost, and field-size independent wafer cost shared by each of the remaining projects from $1/N$ to $1/(N-1)$. Clearly, if none of the assumptions hold, performing a sensitive analysis of shared cost is a difficult task. The shared cost of a project is a function of its chip area and required production volume. It also depends on reticle floorplaning and wafer dicing methods. We obviously need a more sophisticated method for carrying out this analysis task.

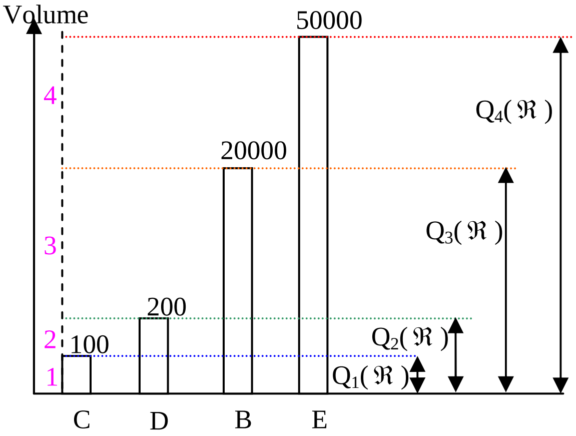


Fig. 2. A mechanism of computing shared cost.

Our sensitivity analysis methodology is built upon our previous work [6] on design space exploration for minimizing MPW production cost. This methodology takes all the chips originally scheduled for an MPW run and performs a design space exploration to find out a minimum cost floorplan. It then does a similar task to obtain a minimum cost floorplan for the case where one chip is removed (added) from/to the MPW run. It calculates the production cost shared by each individual project based on the above two floorplans, respectively. The shared cost change for each project due to removing/adding a project thus can be obtained.

The design space exploration approach employs a reticle floorplanner to arrange chips in a reticle, a wafer dicing algorithm to find wafer dicing plans, and a cost model for computing total MPW fabrication cost and the cost assumed by each individual project. Instead of using the floorplanner proposed in [8], we use a floorplanner based on B*-tree [11] with an objective function to maximize the compatibility among chips with large production volumes [6]. We can vary a weighing factor in the objective function to trade compatibility for reticle size to perform a design space exploration. Two chips are compatible if dicing out one chip would not destroy the other chip and vice versa. This floorplanner is proved to be very effective [12,13]. It outperforms the quadri-section floorplanner [14] by 16~27% in the number of required wafers.

As for wafer dicing, we implement an integer linear programming (ILP) model using the maximal independent sets in a reticle conflict graph as it was done in [14]. This kind of dicing method was first used in [15]. Since it does not work well for low volume production, as suggested in [15] we will use HVMIS-SA-Z for low volume dicing and the ILP model for high volume dicing. Note that the purpose of wafer dicing is to determine the number of required wafers. Once we know the number of wafers we can use (1) to compute the total fabrication cost of an MPW run. We then select a floorplan that results in a minimum production cost.

5: EXPERIMENTAL RESULTS

In this section we perform some experiments to see how sensitive the MPW production cost is to removing a project from an MPW run. Note that the consequence of adding a project is just opposite to that of removing a project. We assume 300 mm wafers are used. We use the mask cost data in Table 1. We also use the data about cost per exposure in [9] to compute the exposure cost per wafer. The cost per exposure is $2.5 for very critical layers, $1.5 for critical layers, and $0.5 for non-critical layers. Field-size independent wafer cost is set to $2500 per wafers [16]. Two test cases in Table 2 [6] are used in our experiments, where $W_{\max}$ and $H_{\max}$ are respectively the maximum reticle width and height. I7 is a test case that combines all projects in the test cases from I1 to I4 in [6]. Each project in a test case is denoted by $N(w,h|1X)$ where N, w, h, and 1X are chip's name, chip width, chip height, and chip production volume.

Table 2. Test cases.

<table>
<tr><th colspan="2">(<i>w, h | 1X required volume</i>) <i>W_{max} =20mm H_{max} =20mm</i></th></tr>
<tr><td>I6</td><td>A(6.5,6.5 | 60), B(4.5,5.0 | 100), C(5.5,1.5 | 120), D(4.5,3.0 | 120), E(6.5,3.5 | 160), F(4.5,3.5 | 160), G(6.5,8.0 | 200), H(3.3,3.5 | 200), I(2.5,3.5 | 200), J(3.5,2.5 | 200), K(7.5,2.5 | 200), L(4.0,2.5 | 200), M(2.5,2.5 | 200)</td></tr>
<tr><td>I7</td><td>A(9.5,9.5 | 60),B(2.0,2.0 | 200),C(2.5,2.5 | 200),D(4.0,5.5 | 80), E(4.0,3.78 | 150),F(3.0,3.0 | 80),G(3.0,2.2 | 80), H(7.0,2.5 | 120), I(5.0,2.0 | 120),J(5.0,3.0 | 120), K(3.0,2.0 | 120), L(2.0,2.0 | 120), M(4.0,3.0 | 60),N(6.5,7.0 | 60),O(2.0,2.5 | 200),P(2.0,1.0 | 200), Q(1.5,2.5 | 400),R(5.0,3.0 | 400),S(2.0,1.5 | 600), T(3.0,2.5 | 600)</td></tr>
</table>

Figs. 3 and 4 show the total MPW fabrication costs prior to and after removing a project where 10X and 100X volumes denote the volume of each project is scaled by 10 times and 100 times, respectively. Since the MPW production cost depends on chip area and production volume, the chip we removed is the one with large size, medium size, small size, large volume, medium volume, and small volume, respectively. For example, we remove G with large size and A with small volume in I6. For I7, we also consider removing a chip like chip T with large volume and with an area larger than S's. "Original" in the X-axis denotes the case where no chip is removed. *Cmask*, *Cexposure*, and *Cwafer* are mask cost, exposure cost, and field-size independent wafer cost, respectively. From these figures, we can see how the total MPW fabrication cost and how the portion of mask tooling cost vary with production volumes. Figs. 5, 6, and 7 give the cost differences for the remaining projects due to removing a project from an MPW run. The chip area and required volume of a project are also printed in the inset of a figure.

For 1X production volume, the mask tooling cost dominates the total MPW production cost. Removing a chip with large area like G in I6 and A in I7 reduces more total production cost, but this does not mean the cost shared by each of the remaining projects would also be reduced. On the contrary, their shared costs are increased. This is due to the increase in mask cost shared among the remaining projects. Interestingly, removing any of the investigated projects from I6 increases the cost shared by each remaining project, whereas removing a project like S or T with large volume but smaller area from I7, the cost shared by each remaining project is decreased. As production volume increases, we can hardly draw any rules from the data given in Fig. 6. When the volume further increases to 100X, wafer fabrication cost becomes more dominating the total MPW production cost, but still no single rule can be drawn to characterize the change of shared cost assumed by each project. The reason for this is that there does not exist a single term in (2) that can completely dominate the total MPW fabrication cost for 10X and 100X production volumes.

The sensitivity analysis is especially useful for a foundry or a design service company to decide whether to accommodate one more project in an MPW run. For example, adding project A or N to I7 with 1X production volume is good for all the projects originally in the MPW run since the cost shared by each of the projects is reduced. On the contrary, it is no good adding project P, S, or T to I7.

6: CONCLUSIONS

This paper has performed a series of experiments on cost sensitivity analysis of removing/adding a project from/to an MPW run. Our analysis shows that removing a project with large chip area would increase the cost shared by each individual project for low-volume production. However, as production volume increases, no single rule can be drawn to characterize the change of the shared cost of each project. We suggest that a full analysis be carried out once the status of an MPW run is changed. The analysis is very useful for a foundry or a design service company to decide whether to accommodate one more project in an MPW run.

REFERENCES

[1] M. LaPedus. Is IC industry heading to the $10 million photomask?. Semiconductor Business News, Oct. 7, 2002.

[2] L. Pillegi, H. Schmit, A. J. Strojwas, P. Gopalakrishnan, V. Kheterpal, A. Koorapaty, C. Patel, V. Rovner, and K. Y. Tong, "Exploring regular fabrics to optimize the performance-cost trade-off," DAC, pp. 782-787, June 2003.

[3] C. A. Pina, "MOSIS: IC prototyping and low volume production service," Proc. of Intl. Conf. on Microelectronic Systems Education, 2001.

[4] A. B. Kahng, I. Mandoiu, Q. Wang, X. Xu, and A. Z. Zelikovsky, "Multi-project reticle floorplanning and wafer dicing," Proc. of ISPD, pp.70-77, 2004.

[5] G. Xu, R. Tian, D. Z. Pan, D.F. Wong, "A multi-objective floorplanner for shuttle mask," Proc. of SPIE, Vol 5567, pp. 340-350, 2004.

[6] M. C. Wu and R. B. Lin, "Reticle ploorplanning and wafer dicing for multiple project wafers," ISQED, pp. 610-615, 2005.

[7] M. C. Wu and R. B. Lin, "Multiple project wafers for medium-volume IC production," ISCAS, pp. 4725-4728, 2005.

[8] R. B. Lin , M. C. Wu, W. C. Tseng, M. H. Kuo, T. Y. Lin, and S. C. Tasi, "Design space exploration for minimizing multi-project wafer production cost," ASPDAC, pp. 783-788, 2006.

[9] D. Pramanik, H. Kamberian, C. Progler, M. Sanie, and D. Pinto, "Cost effective strategies for ASIC masks," Proc. Of SPIE, vol. 5043, pp. 142-152, 2003.

[10] S. Miraglia, C. Blouin, G. Boldman, S. Judd, T. Richardson, and D. Yao, "ABC modeling: advanced features," Advanced Semiconductor Manufacturing Conference, pp. 336-339, 2002.

[11] Y. C. Chang, Y. W. Chang, G. M. Wu, and S. W. Wu, "B*-tree: a new representation for non-slicing floorplans," DAC, pp. 458-463, 2000.

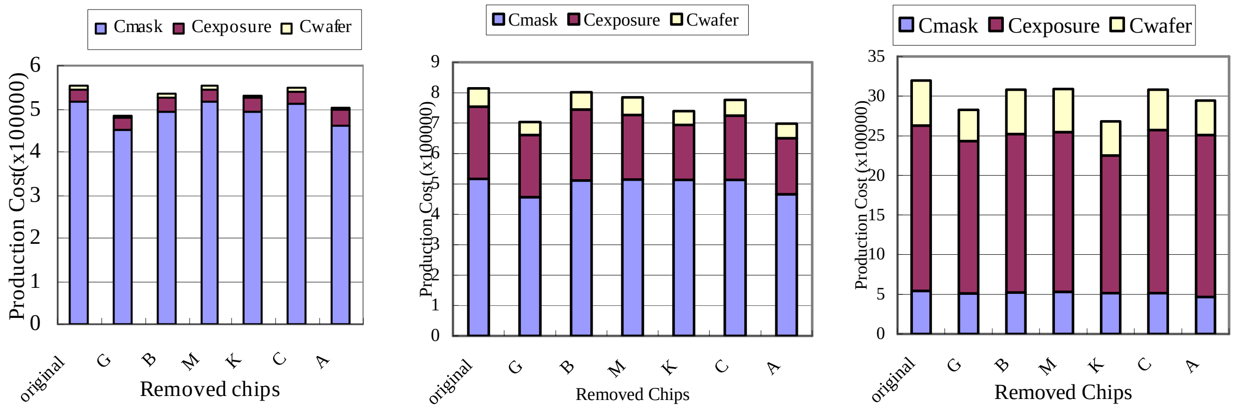
[12] R. B. Lin, M. C. Wu, and S. C. Tsai, "Reticle design for minimizing multi-project wafer production cost," submitted to IEEE T-ASE.

[13] M. C. Wu, S. C. Tsai, and R. B. Lin, "Floorplanning multiple reticles for multi-project wafers," VLSI-DAT, 2006.

[14] A. B. Kahng, I.I. Mandoiu, X. Xu, and A. Zelikovsky, "Yield-driven multi-project reticle design and wafer dicing," Proc. of 25th Annual BACUS Symposium on Photomask Technology, pp. 1247-1257, 2005.

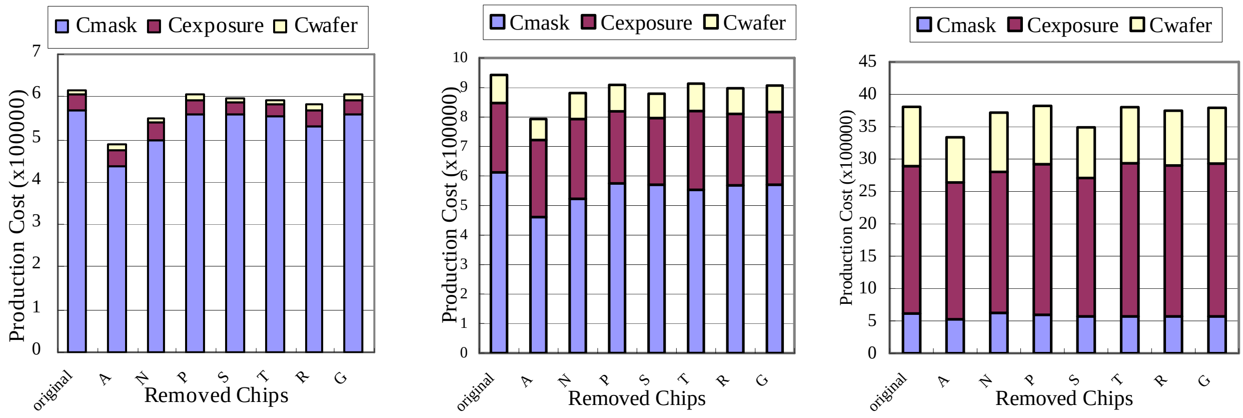
[15] M. C. Wu and R. B. Lin, "A comparative study on dicing of multiple project wafers," IEEE Computer Society Annual Symposium on VLSI, pp. 314-315, 2005.

[16] "Optical Lithography Cost of Ownership (COO) – Final Report for LITG501," International SEMATECH.



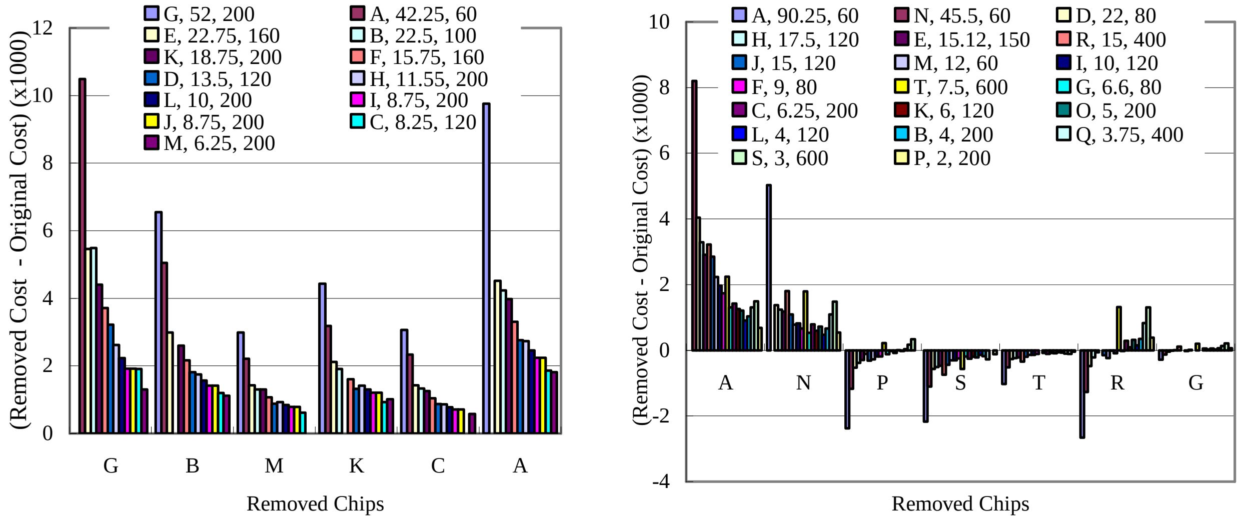
(a) 1X volume (b) 10X volume (c) 100X volume

Fig. 3. MPW production cost for I6.



(a) 1X volume (b) 10X volume (c) 100X volume

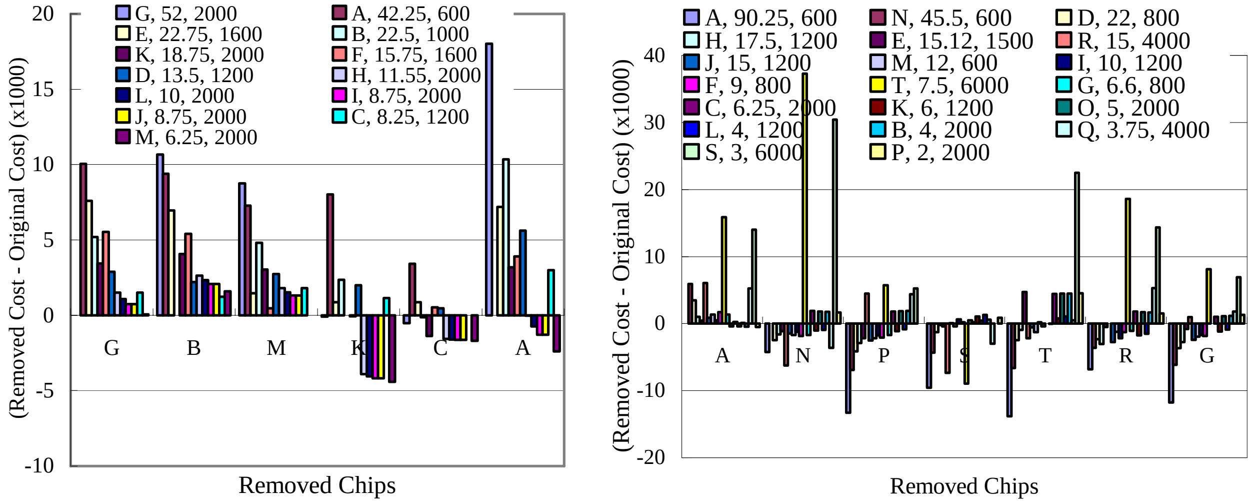
Fig. 4. MPW production cost for I7.



(a) I6 with 1X volume

(b) I7 with 1X volume

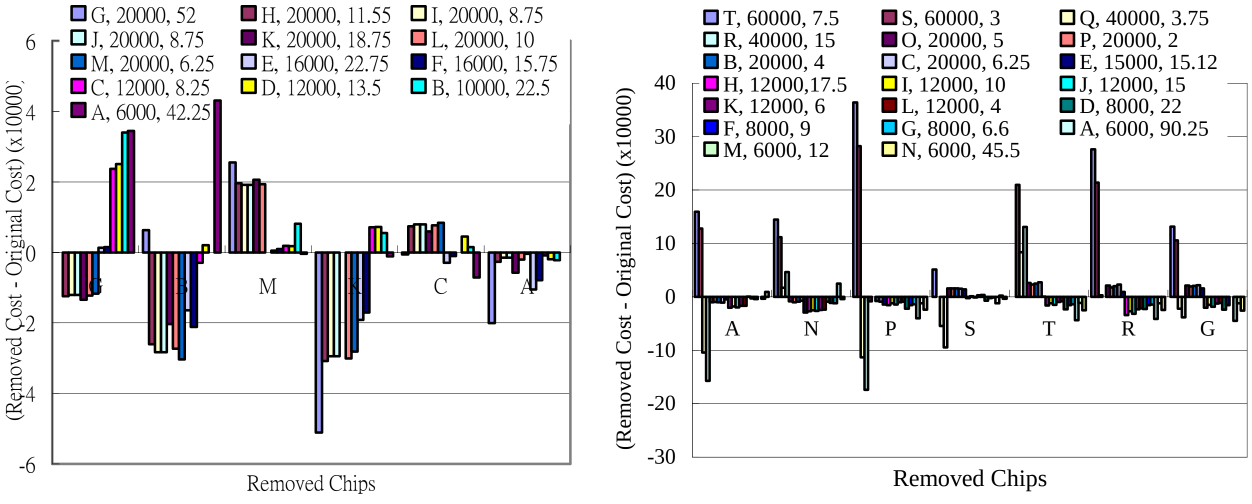
Fig. 5. Shared cost changes due to removing projects form I6 and I7 with 1X volume.



(a) I6 with 10X volume

(b) I7 with 10X volume

Fig. 6. Shared cost changes due to removing projects form I6 and I7 with 10X volume.



(a) I6 with 100X volume

(b) I7 with 100X volume

Fig. 7. Shared cost changes due to removing projects form I6 and I7 with 100X volume.