Transform designs to chips, an end user point of view on mask making

John Y. Chen
NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050, USA

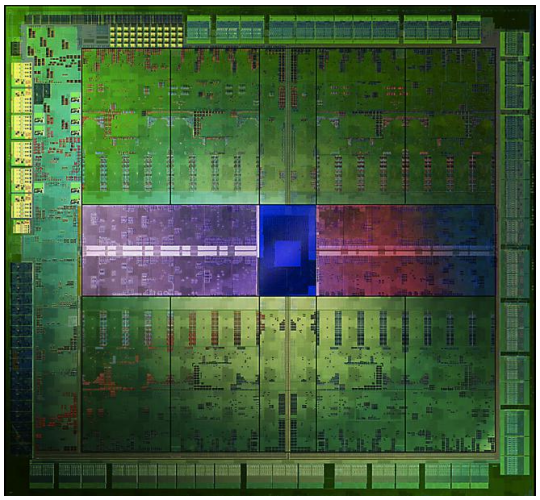
ABSTRACT

Mask is a tool required to replicate a set of complicated IC geometries numerous times producing chips in large volume. It is absolutely crucial to achieve high quality mask in accuracy and perfection. This paper focuses on technology needs which challenge mask-making capabilities including data preparation and Optical Proximity Correction (OPC), Critical Dimension (CD) / Line Edge Roughness (LER) control, alignment and defect elimination. The impact of lithography and mask making on the design and manufacturing of new products is discussed from an end user perspective. The paper emphasizes *performance, precision and perfection* and the necessity of the three *p's* for the continuation of Moore's law.

Keywords: Performance, Precision, Perfection, Perf/watt, OPC, DFM, CD, Alignment

1. INTRODUCTION

Mask making is a key element in a new product introduction cycle, and arguably the most critical part in the manufacturing process due to its huge impact if not done right. Today, the definition of right means perfection. Whether it is data completeness or pattern fidelity (placement and size), the requirement is zero flaw. Moreover, mask making including OPC, serves as a bridge which carries design data to manufacturing and feeds back OPC-unfriendly layouts to designers for better Design for Manufacturing (DFM). This process has worked well in the past and allowed end users in fabless as well as IDM companies to build most complicated chips in the world.



For example, shown in Fig.1 is NVIDIA's latest Kepler GPU (Graphics Processor Unit) chip manufactured using 28nm technology. It consists of 3.5 billion transistors, 10 billion contacts and 12 billion vias connecting 10 interconnect layers. We are now making test chips for future technology nodes and design rules. However, we are facing new challenges and would encounter even more as we follow Moore's law to build future chips with higher performance, less power dissipation and more functions.

There are basically **three** major technological challenges, namely **Performance, Precision and Perfection** and they all critically depend on mask-making capabilities including data preparation/OPC, CD control, alignment, and defect elimination. In the following sections, we will discuss these challenges and possible solutions.

Fig. 1 Kepler GK104 in GTX680 graphics card.

2. CHALLENGES

Performance, in fact **efficient performance** referred as Perf/watt is the most important figure of merit for chips made today. We define Perf in terms of GFLOPS (Giga Floating point Operations per Second) and it is proportional to the product of transistor count and circuit speed. Fig.2 shows the continuous improvement of chip performance as the technology advances[1].

Photomask Technology 2012, edited by Frank E. Abboud, Thomas B. Faure, Proc. of SPIE Vol. 8522, 852202 · © 2012 SPIE · CCC code: 0277-786/12/$18 · doi: 10.1117/12.981267

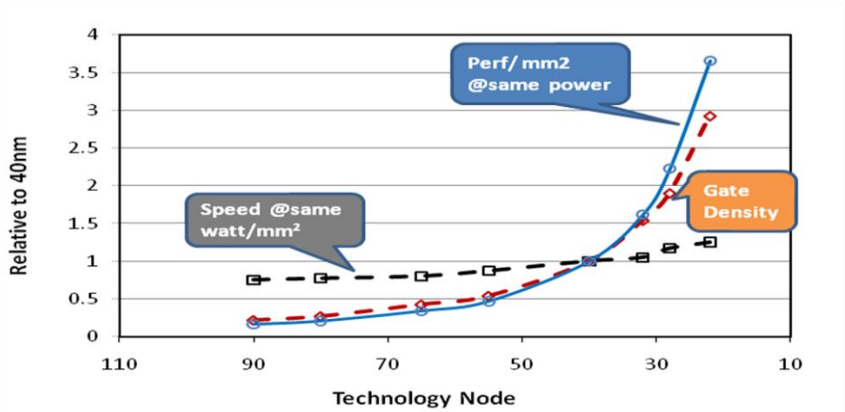
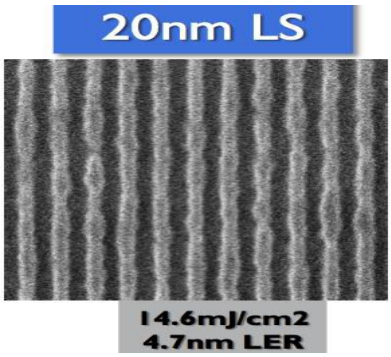


Fig.2 Scaling GPU performance with technology [1], Perf = # of GFLOPS *(Ref.1)*

Notice that under the power constraint, most Perf improvement comes from doubling the transistor density based on Moore's law [2]. This is done with ever shrinking design rules that have been achieved by the ability of replicating finer lines and spaces through higher resolution masks. Meanwhile because transistor count has been growing, the amount of data and hardware needed for making the mask has been going up tremendously. Not only are we challenged to print ever shrinking tiny geometries, but also to print more than billions of them for each masking layer.

Precision in dimension and placement is another challenge that we are facing. Every nanometer counts because our transistors are approaching 20nm in length. For transistors at this dimension, their electrical behavior changes a lot for just 1nm variation. If a 20nm-long transistor becomes 21nm, it can lose 5% in speed when it's ON, but if it becomes 19nm, it can leak as much as 2x more current when it's OFF. This is why the minimum physical gate length hasn't been scaling per Moore's law in the recent technology nodes. Designers today can purposely bias the length of the transistor by a couple of nanometers to create another type of transistors for either higher speed or low leakage. However, if the CD cannot be controlled precisely, the two types of transistors can hardly be distinguished.

LER shown in Fig. 3 is another problem as the line width is being further reduced. Even 1nm LER per side of a 20nm-wide line represents a +–10% CD variation along the line. This is particularly problematic when associated with the gate of a transistor for which it's electrically equivalent to a bunch of sub-transistors with varying gate lengths connected in parallel.



- A X'tor with LER = many X'tors with varying *L's* connected in parallel
- Longer *L* X'tors gives slower speed
- Shorter *L* X'tors gives
  a. lower Vt,
  b. much higher (exponentially proportional to 1/*L*) Ioff
  c. Higher Miller cap (Cgd)

Fig. 3 Line Edge Roughness (LER) and effects on a transistor gate [3]. *(Ref. 3)*

While the ones with longer than designed gate slow down the speed, the shorter ones generate a lot more leakages, and this effect is difficult to model resulting in unpredictable circuit behavior.

The placement of the geometries is more critical today than ever because of the narrower spaces between the features in the same masking layer and in the adjacent masking layers. Fig.4 is an example showing the narrow spacing between a

Via 1 (V1) at the end of a Metal 1 (M1) line and another adjacent M1 line must be precisely controlled to allow sufficient space. Otherwise, a short may occur. Even with a space that a short doesn't occur, if it's too narrow, it may lead dielectric breakdown over time.

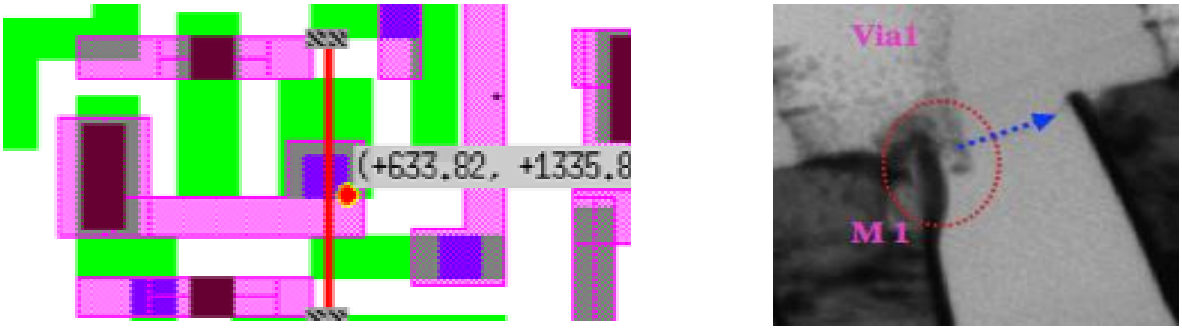


Fig. 4 Layout and SEM cross section[4] showing precision control of the narrow space between Via 1 (V1) and Metal 1 (M1). V1 is the via connecting M1 to M2 (Metal 2). ------> shows the spacing in discussion *(Ref. 4)*

A short kills the die, an almost short is even worse because it manifests itself as a TDDB (Time Dependent Dielectric Breakdown) related reliability problem [4]. In order to have the right spacing between the V1 and the adjacent M1, one must control V1 size and M1 space tightly, and align the V1 layer to M1 layer precisely.

Perfection is here defined as the degree of flawless. Imperfection is quantified as DPPM or DPPB (Defective Parts Per Million or Per Billion). For a chip with billions contacts or vias, one must have the corresponding defect level less than 1 DPPB, near perfection. Even designers apply DFM such as redundant vias and rectangular vias to mitigate the problem, a large number of single minimum-size vias in the range of few hundreds millions is often left on the chip due to layout constraints to keep a reasonably competitive die size. Let's refer these vias to worst-case vias meaning that they are least DFM friendly though complying to design rules. Using the example shown in Fig.4 with some typical numbers for illustration purpose, one can see how much precision and perfection are required.

With 100 million worst-case vias on a 20nm chip, if 10nm is the minimum spacing required to guarantee product lifetime reliability, chip failure would be about 240 DPPM if the nth via (VIAn) and nth metal (Mn) layer CD's are controlled at 2.0nm and 1.4nm respectively, and the VIAn-to-Mn alignment is as tight as 2.3nm. Sensitivity analyses show that a sub-nm difference in the sigma of Mn CD, VIAn CD, or misalignment distribution can alter the product lifetime drastically. Fig. 5 is an example illustrating that just one tenth of a nanometer less in this critical overly control would increase the chip failure rate by as much as an order of magnitude!

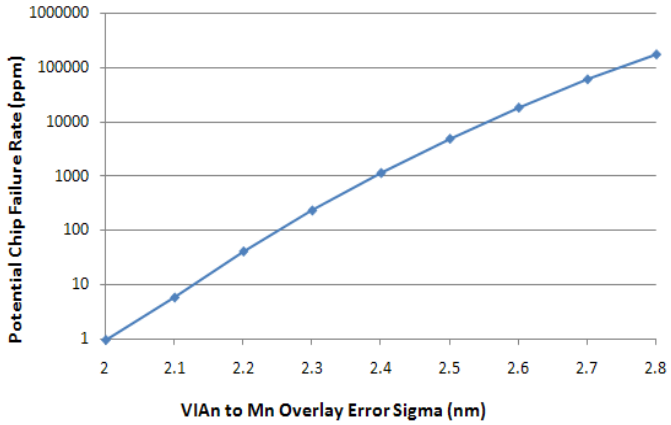


Fig. 5 Potential Chip Failure Rate in DPPM vs. VIAn-to-Mn misalignment sigma (σ) in nm for a chip with 100M worst-case Via's made by 20nm technology. The calculation assumes the VIAn CD and Mn CD 1σ variation are 2.0nm and 1.4nm respectively.

This is not so shocking if one realizes that we are now dealing with 100 million vias per chip and a slight change in the sigma of the misalignment distribution can affect the tail of the distribution greatly.

As the number of vias increases, the precision requirement is more stringent and the defect needs to be almost zero. In addition to the precision of the CD and the overly control, the cleanliness on either mask as well as on other parts of the manufacturing process must approach perfection. Again, considering mask (193 immersion or EUV mask) is tooling which gets replicated millions times to produce silicon chips, its quality in CD, alignment, and defect control must be perfect.

3. CONCLUSION

In conclusion, as an end user, I admire what our lithography experts and mask makers have accomplished so far. Looking into the challenges ahead, considering nanometer resolution, complicated patterns and numerous geometries, I can't help preaching again for **Performance, Precision and Perfection**. If there is anything else to add, that is **pricing,** pricing based on cost reduction by faster CPU/GPU data manipulation [5], more effective and efficient OPC, and shorter e-beam writing time. The three **P's** and improved pricing would help users continue building better products with lower transistor cost for many generations to come.

REFERENCES

1. J.Y. Chen, "GPU Technology Trends and Future Requirements," IEDM Paper 1.1, Baltimore, Maryland 2009.
2. G. Moore, "Cramming More Components onto Integrated Circuits," Electronics, 38(8), April 1965.
3. G. Vandenberghe, "EUV Lithography, what are the major technical hurdles other than source intensity?" IMEC presentation at NVIDIA, Santa Clara, CA, 2011.
4. W. Liu, Y. K. Lim, J. B. Tan, W. Y. Zhang, H. Liu, and S. Y. Siah, "Study of TDDB Reliability in Misaligned Via Chain Structures", IRPS Paper 3A.4.1, Anaheim, California, 2012.
5. CUDA Technology, NVIDIA, 2012, http://www.nvidia.com/CUDA.