The Road Ahead

Bringing down NRE

Andrew B. Kahng
University of California, San Diego

In my July-August 2002 column ("The cost of design," pp. 136, 135), I highlighted today's staggering levels of design nonrecurring engineering (NRE) costs, which routinely reach tens of millions of dollars. But manufacturing NRE, the cost of a mask set and probe card, will reach $1 million at the 90-nm technology node. This cost is staggering as well, and puts a significant damper on semiconductor-based innovation. As the semiconductor industry deploys optical proximity corrections, phase-shifting masks, and area fill to better control subwavelength lithography, will the million-dollar mask set become a fact of life, or can the industry reduce this key component of NRE?

Shape complexity and mask cost

Mask costs rise because of shape complexity. Adding such proximity corrections as serifs, hammerheads, and scattering bars transforms a simple rectangle in the designer's layout viewer into several disjoint polygons (having 30 or more corners) on the mask. Layout rules and rule checking become exceedingly complex to accommodate such possibilities. According to the 2001 *International Technology Roadmap for Semiconductors*, the mask data volume for a single critical layer will reach 216 Gbytes at the 90-nm node, up from 64 Gbytes at 130 nm. This explosive growth rate, given that transistor count only doubles between successive nodes, reveals the increase in mask shape complexity as optical lithography moves ever deeper into the subwavelength regime. Increased data volume also results from the need to introduce dummy shapes that make layouts more uniform to reduce process variability.

To simplify matters, let's ignore two second-order and somewhat countervailing factors: mask inspection (whether you can check what you write on a mask, and whether yield loss in mask production occurs because of imperfect inspection capability), and cost-volume dynamics (whether costs will drop substantially when production volumes increase).

Today's mask flow takes a GDSII file (GDSII is a format for transferring 2D graphical-design data) from the designer's layout tool, fractures it into MEBES (manufacturing electron-beam exposure system) or other proprietary data formats that represent primitive mask-writer operations (flashes), and then passes the MEBES data to mask-writing tools. The cost of the mask is essentially proportional to write time, which is proportional to the number of flashes. Reducing the number of flashes will increase production throughput and proportionally reduce mask costs.

With today's high-voltage (50 kV), vector, shaped e-beam tools, the mask-writing process has many knobs for optimization:

- Multiple writing passes add fault tolerance but reduce throughput.
- Higher e-beam current density increases throughput but reduces pattern fidelity because of photoresist heating and distortion.
- Larger areas of individual flashes increase throughput but decrease patterning accuracy.
- Newer writers with 45° apertures provide a faster alternative to approximating an angled line edge with a series of tiny rectangles, but angled lines still require long write times because they must be decomposed into triangular flashes.

Not enough work to date has explored these degrees of freedom in optimizing the mask-writing process.

0740-7475/03/$17.00 © 2003 IEEE Copublished by the IEEE CS and the IEEE CASS

Design technology for reduced-mask NRE

Recent design-for-manufacturing initiatives in EDA address cost issues from a manufacturing-yield standpoint. However, these efforts focus on the incremental costs of manufacturing functional die. Opportunities remain for new technologies that attack fixed, nonrecurring costs. Here are three promising directions.

Parametric yield-driven lithographic corrections

Today, when optical proximity or other lithographic corrections are applied, there is no concept of incremental "dollars per wafer" return on investment, or even of impact on design functionality. Yet, it makes little sense to blindly impose (as is common today) equal pattern fidelity for a critical gate, a noncritical poly wire, and a piece of the company logo. Recent research suggests that substantial reductions in mask layout complexity are possible without any loss of parametric yield. These design optimizations are possible based on knowledge of slacks, sensitivities, and criticalities that pass from the design flow. Supporting technologies include statistical static timing and noise analyses, and variability models for lithographic corrections.

Mask-writer throughput improvements

Depending on which tool you use to fracture a GDSII layout into flashes for the mask writer, up to 20× differences in write times are possible. Improved, writer-specific fracturing is clearly the most direct way to improve mask throughput. Again, knowledge of design function and parametric yield sensitivities will let designers choose each shape from a cost-accuracy tradeoff continuum. Complementary optimizations could tune beam currents and the detailed scheduling of flashes to maximize writer throughput while observing the allowable limits of photoresist heating and distortion.

Amortizations

Finally, amortizing various NRE costs could enable another class of NRE reductions. For low-volume and early prototyping contexts, a dedicated mask set and wafer lot comprise a financial showstopper. Yet, production could be cost-feasible if multiple designs share the masks and wafers. Indeed, the concept of a multiproject wafer, or shuttle, is now well-established in both academia and industry. New tools at the design-manufacturing interface, coupled with new production technologies such as waterjet-guided lasers for flexible dicing, will allow the concept of amortization to generalize along multiple axes.

Cost reductions are also possible through restricted layout styles that fit with high-volume, low-cost, premanufactured mask blanks. An excellent example is the Phase Phirst! proposal by Levenson et al.[1] Although layout restrictions imply a loss of layout density, I view this as a "soft reset" that is far more survivable than any cost-induced "hard reset" that would stop Moore's law and the semiconductor retooling cycle.

THE PROBLEM of manufacturing NRE is, while daunting, absolutely within the grasp of design technologists. The technology directions I've just described give the EDA R&D community excellent opportunities to deliver tangible value to semiconductor companies by reducing manufacturing NRE and expanding the universe of cost-feasible designs. ■

Reference

1. M.D. Levenson et al., "Phase Phirst! An Improved Strong-PSM Paradigm," *Proc. 20th Ann. BACUS Symp. Photomask Technology*, SPIE, 2000, vol. 4186, pp. 395-404.

■ Direct questions and comments about this department to Andrew B. Kahng, University of California, San Diego, Computer Science and Engineering Dept., 9500 Gilman Dr., MC 0114, La Jolla, CA 92093-0114; abk@ucsd.edu.