# Application of the Cost-per-Good-Die Metric for Process Design Co-optimization

Tejas Jhaveri[1,2], Umut Arslan[1], Vyacheslav Rovner[1,2], Andrzej Strojwas[1,2] & Larry Pileggi[1,2]
[1]Carnegie Mellon University, Pittsburgh, PA 15213
[2]PDF Solutions, 5830 Ellsworth Ave, Suite 304, Pittsburgh, PA 15232

## ABSTRACT

The semiconductor industry has pursued a rapid pace of technology scaling to achieve an exponential component cost reduction. Over the years the goal of technology scaling has been distilled down to two discrete targets. Process engineers focus on sustaining wafer costs, while manufacturing smaller dimensions whereas design engineers work towards creating newer IC designs that can feed the next generation of electronic products. In doing so, the impact of process choices made by the manufacturing community on the design of ICs and vice-versa were conveniently ignored. However, with the lack of cost effective lithography solutions at the forefront, the process and design communities are struggling to minimize IC die costs by following the described traditional scaling practices. In this paper we discuss a framework for quantifying the economic impact of design and process decisions on the overall product by comparing the cost-per-good-die. We discuss the intricacies involved in computing the cost-per-good-die as we make design and technology choices. We also discuss the impact of design and lithography choices for the 32nm and 22nm technology node. The results demonstrate a strong volume dependence on the optimum design style and corresponding lithography strategy. Most importantly, using this framework process and design engineers can collaborate to define design styles and lithography solutions that will lead to continued IC cost scaling.

**Keywords:** Application specific integrated circuits, CMOS integrated circuits, Lithography, Integrated circuit economics

## 1. INTRODUCTION

The semiconductor industry has been celebrated for its technical prowess, marked by its superior rate of technology development and commercialization. The impact of these breakthroughs on our society is irrefutable. However, the adoption and commercialization of any technology, regardless of it technical superiority, is impossible without a tangible economic justification. Technologies that show significant economical advantages are adopted aggressively, whereas the adoption of technologies that cannot be economically commercialized have been pushed out till theyare deemed to be economically viable. Similarly the economics have been a driving factor behind the industries explosive growth. To shed more light on the growth factors, we begin by providing an historic overview of technology scaling and an indication of how it must evolve to drive continued growth of the industry.

### 1.1. Dynamics of Technology Scaling

In 1965, Gordon Moore was the first to outline an exponential growth for the industry as long as engineers continued to increase the number of components (transistors) on an IC [1]. Albeit, Moore altered the magnitude of his predictions in 1975, the fundamental economical justifications have remained unchanged even to this day. As per his extrapolations in 1975, the number of transistors must be doubled every two years in order to minimize the cost per transistor [2]. In the years to follow engineers across the industry have worked persistently towards doubling the number of transistors on a fixed silicon real estate by aggressively scaling resolution by 70% and hence doubling the component density on ICs every two years. A closer examination of the 1965 paper suggests that the exponential increase in the complexity of ICs is only viable as long as we can sustain high yields without significantly increasing processing costs. Unfortunately, both of these assumptions are no longer true. It is quite ironic that in a persistent effort to double transistor density to meet Moore's Law, the industry, as a whole, has failed to minimize the cost-per-component and has desecrated the spirit of Moore's observations.

The semiconductor industry has also diversified significantly since Moore made his initial predictions. At present, a wider range of ICs with varying product volumes, ranging from few thousand dies to 100s of millions of dies are

Design for Manufacturability through Design-Process Integration IV, edited by Michael L. Rieger, Joerg Thiele,
Proc. of SPIE Vol. 7641, 764104 · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.846556

produced to support the various products being developed by the electronics industry. As the non-recurring costs per ICs continue to grow, the effective cost per component tends to increase dramatically with decreasing product volumes, and as such have made it very challenging to extend the economical benefits of scaling for low to medium volume products. It is apparent that tenaciously following the path of doubling component density every generation is not sufficient moving forward.

### 1.2. Outline & Scope

Over three decades after Moore's revised predictions it is critical that we determine strategies and technologies that will help us minimize the cost per component [3]-[5]. In section II of this paper we provide an overview of the metrics used for technology scaling and discuss the value of using the cost-per-good-die metric. Section III describes our framework for computing the cost-per-good-die. Section IV provides the results and a discussion of applying the cost per good die metric for 32nm logic design whereas Section V applies the cost per good die metric for 22nm node technology definition.

## 2. METRICS FOR TECHNOLOGY SCALING

Density (or die area) has been the metric most widely adopted to gauge the effectiveness of technology scaling, ever since Moore made his profound observations. These observations were made while the industry was in its adolescence; at the time when IC manufacturing trends allowed Moore to claim that the cost per unit area of silicon and product yield could be sufficiently scaled while doubling component density every generation. More recently the challenges associated with achieving sufficient yields for sub-100nm products has forced enlightened engineers to compromise on the strict density scaling requirements [3]-[5].

The focus has shifted from a purely area based scaling strategy, to a strategy where design is scaled only to the extent that the number of good-dies-per-wafer ($N_{gdw}$) is maximized. The right trade-offs between the number of chips on a wafer ($N_{dw}$) and its manufacturability or yield (Y) is sought.

$$N_{gdw} = N_{dw} Y \tag{1}$$

More recently, it is experienced that maximizing the number of good-dies-per-wafer is not sufficient. The impending challenge with scaling the wavelength of light source used for optical lithography has led to the adoption of alternate lithography techniques such as Source Mask Optimization (SMO) [24], Double Patterning Technologies (DPT) [17]-[20] and Interference Assisted Lithography (Intf) [25, 26]. The escalating process complexity and costs require a comprehensive evaluation of the improvement in density and yield enabled by increasing the cost per unit silicon. To achieve this end goal we need a different metric to gauge the success of technology scaling.

The cost per transistor has been suggested in the past [3]-[5]. Although a physically fundamental metric, it lacks practical significance in the era of nanometer System on Chips (SOCs). The cost per transistor tends to vary significantly among different types of IC sub-components. Modern SOCs are designed with varying amounts of logic, memory, analog and I/O components along with other non-transistor devices for RF and MEMs applications that are integrated onto these SOCs. This makes it difficult to relate the cost per transistor to the true component costs. To focus on minimizing the component costs we recommend the use of the cost-per-good-die ($C_{gd}$) as the metric to evaluate the success of technology scaling. It must be pointed out that the cost-per-good-die is only effective when used to compare costs between similar designs.

$$C_{gd} = \frac{C_w}{N_{gdw}} \tag{2}$$

The cost-per-good-die is determined as the ratio of the cost-per-wafer ($C_w$) and the number of good dies per wafer ($N_{gdw}$) [6]. The cost-per-wafer includes the recurring and nonrecurring costs associated to the engineering resources, materials, software, and equipment utilized both during the design and manufacturing processes. It is intuitive that the focus for IC cost minimization is not only to maximize the number of good dies per wafer but also to simultaneously minimize the wafer cost. Using this metric it is critical that we scale IC products only to the extent that economies of scale can be sustained.

## 3. FRAMEWORK FOR COMPUTING THE COST-PER-GOOD-DIE

The primary elements for the cost-per-good-die computation are the product yield, number of dies-per-wafer, and the cost-per-wafer as have been discussed in the earlier section. As discussed in earlier publications [9, 14] the number-of-dies is computed based on product dimensions, yield is computed by approximating random, systematic and parametric yield components, and the cost-per-wafer determined by computing the lithography and non-lithography related costs. In this paper, we describe the overall flow of data used in our framework for computing the cost-per-good-die (Figure 1).

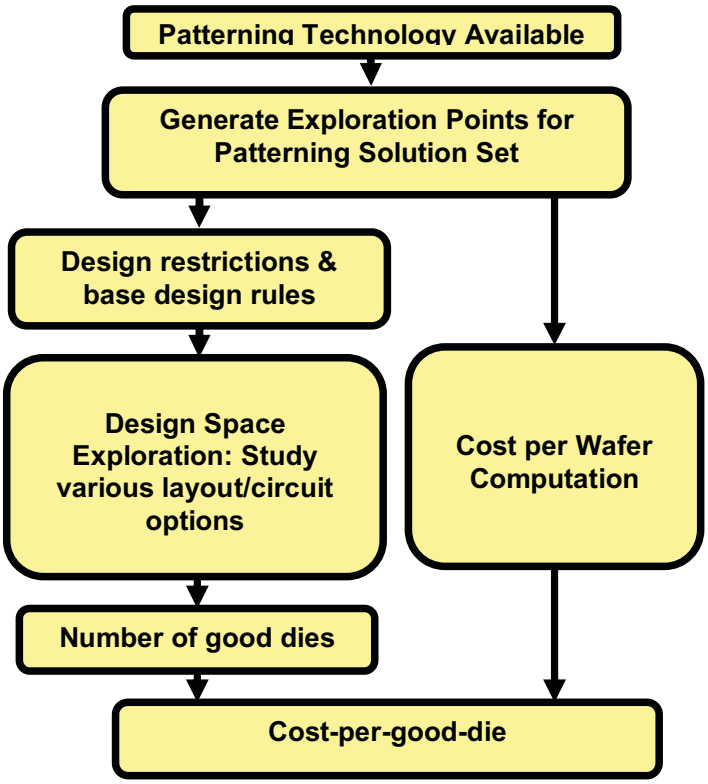



**Figure 1: Framework for cost-per-good-die computation**

At the foremost we begin by evaluating the viable options for a given technology node. The viability of these options is determined based on the availability of tools and supporting infrastructure at the point of technology node / product insertion into volume manufacturing. It is noted that at present not all processing layers use the same lithography option. Hence we consider several combinations of these lithography options for each layer. E.g. for sub-50nm processes the use of DPT is considered for only select layers such as Poly and Contact. Based on these patterning solution sets we determine the design rule restrictions required as well as compute the cost-per-wafer [9, 14]. These design restrictions and design rules are used for design space exploration where various layout/circuit options are considered to determine the solutions that maximize the number-of-good-dies. The number-of-good-dies and cost-per-wafer are used

### 3.1. Design Space Exploration Example: 45nm Ultra Regular SRAM

Integrated systems are embedding growing amounts of SRAM and move from logic-dominant to memory-dominant systems. Consequently, the overall chip area, power and yield are becoming increasingly dominated by SRAMs. To realize efficient and cost-effective systems, SRAM circuits, particularly the bitcell, must be optimally designed for any given patterning solution (i.e. fabric). Thus, we have constructed a framework to explore the bitcell design space and determine the solution that maximizes the number-of-good-dies. The framework essentially consists of area and yield models, and a methodology for design space exploration.

Regular-patterned layouts greatly facilitate area modeling and thus enable exploration of various bitcell topologies. We created a set of parametric models for possible layout configurations of candidate bitcell topologies. For any given

patterning restrictions (i.e. design rules) and parameter values, the configuration that minimizes the bitcell area can be automatically determined. The bitcells use minimum or near-minimum dimensions for maximum density, thereby making them highly susceptible to within-die random variations (e.g. random dopant fluctuations). The within-die variability results in parametric yield loss for SRAMs via several failure mechanisms. We consider failures due to cell flip during read, unsuccessful writes and access-time failures due to increase in the bitline delay. We adopt the definitions proposed in [33] and [34] to measure the read and write noise margins, respectively. The bitline model has a parameterized length including the wire-load estimated using the area model. Each bitcell has to satisfy an extremely high yield requirement to provide acceptable die yield. Due to the requirement of very large sample sizes, the Monte-Carlo (MC) method is too expensive to be applied for analyzing these parametric failures that occur very rarely. Nonetheless, increasing process variability causes linear models to result in large errors, especially while modeling rare events. Therefore, we use a novel piecewise-linear response surface method (PWL-RSM) combined with importance sampling (IS) to analyze SRAM parametric failures [35]. The PWL-RSM adaptively partitions the variation space until sufficient model accuracy is achieved with linear models in local spaces. Next, IS is used to obtain the failure probability with a reduced sample size. This method provides >10K speed-up compared to traditional simulation-based MC, thereby making fast design space exploration feasible.

The bitcell design framework is illustrated in Figure 2. First, the fabric, bitcell topology, required performance, total memory size and basic architectural features (e.g. number of cells per bitline) are specified as inputs. Then, a number of design points are generated using a set of pattern-dependent design parameters (e.g. diffusion width). Layout dimensions and yield are estimated for each design point (only parametric yield is considered in the following example). Next, designs that minimize the failure probability at any given area (i.e. pareto-optimal designs) are determined. Die dimensions and yield is computed for each pareto-optimal design based on the specified memory size and block configuration. Finally, the bitcell design that maximizes the number of good dies is determined as the optimum solution.

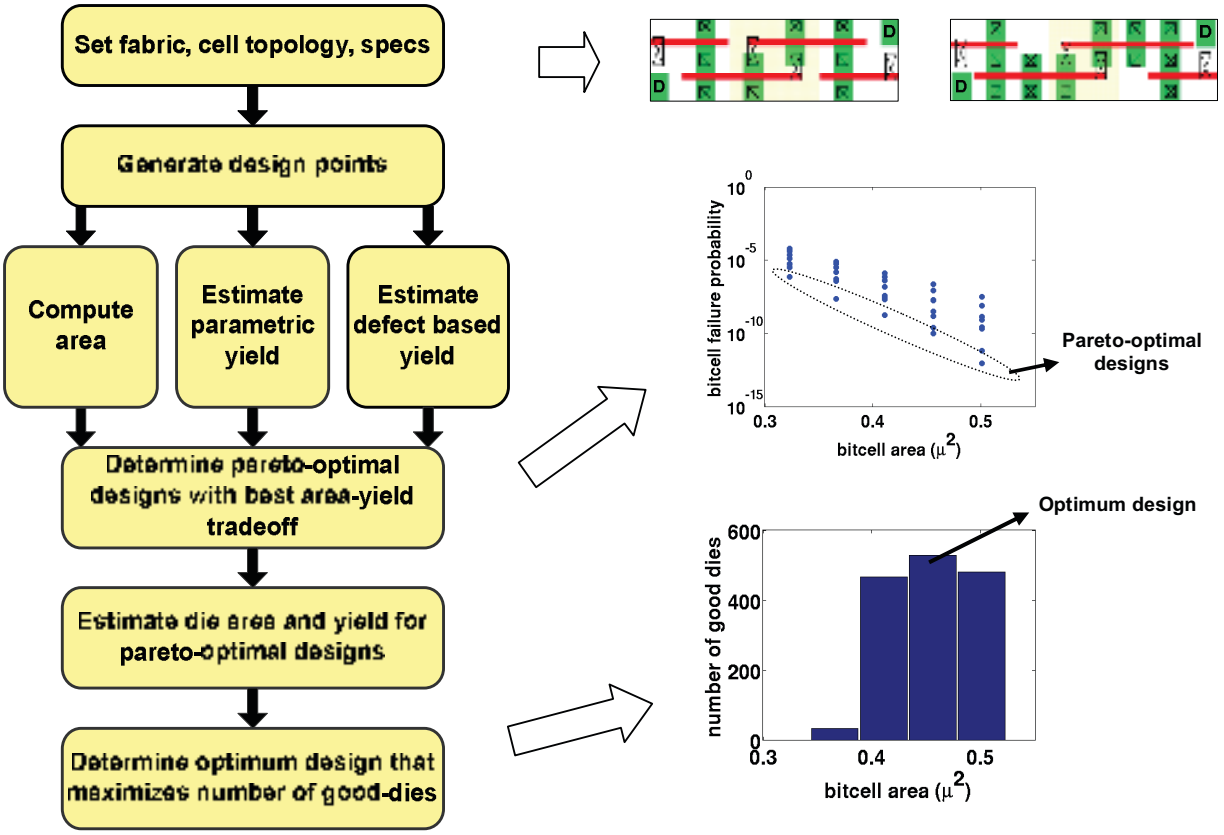



**Figure 2: Framework for SRAM bitcell design**

We have used the framework to design 6T (six transistor) bitcells at 45nm using an extremely-regular fabric based on gratings. The chosen fabric restricts all IC layers to be constructed from gratings except contacts (Off-grid and rectangular contacts provide significant area reduction as will be shown in Table 5). The grating-based diffusion pattern compels 6T bitcell to only use β ratios (i.e. access width/pull-down width) of integer value affecting noise margins and hence parametric yield. In this experiment, we have designed two bitcells having β ratios of 1 and 2. To minimize the failure probability, the β ratios are electrically adjusted by optimizing the wordline voltage as proposed in [36]. In addition, supply voltage is dynamically reduced by 25% during write to improve write-ability at a reduced wordline level. The design parameters include the diffusion width and device threshold voltages (regular or high Vt). The bitline delay is specified to be less than 8 FO4 (fan-out of four) inverter delays for a short-bitline full-swing architecture using 8 cells per column. The device length ($L_{poly}$) is set as 44nm and 48nm for high-performance and low-leakage applications, respectively. The die is assumed to contain 16MB of SRAM composed of 512Kb blocks, each with 1024 rows and 512 columns, arranged in 16x16. Array efficiency for each macro (total bitcell area/total area) is estimated to be 50%. The results from this experiment show that the 6T-β=1 cell provides 24% ($L_{poly}$=44nm) and 6% ($L_{poly}$=48nm) larger number-of-good-dies than the 6T-β=2 topology. Thus, we choose to use the 6T-β1 for the defect-based yield study at 22nm that will be described in Section 5.

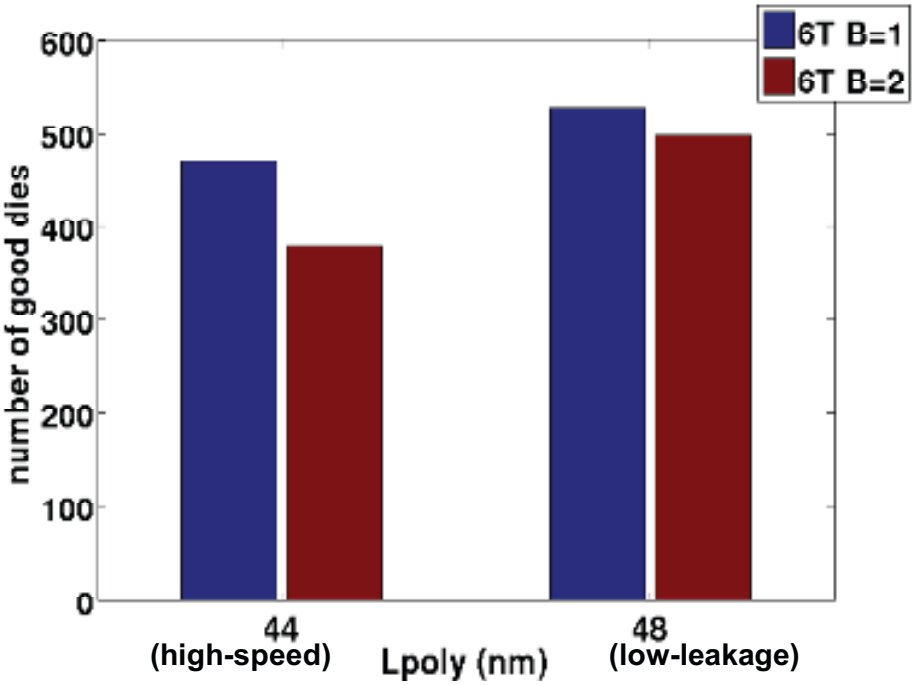
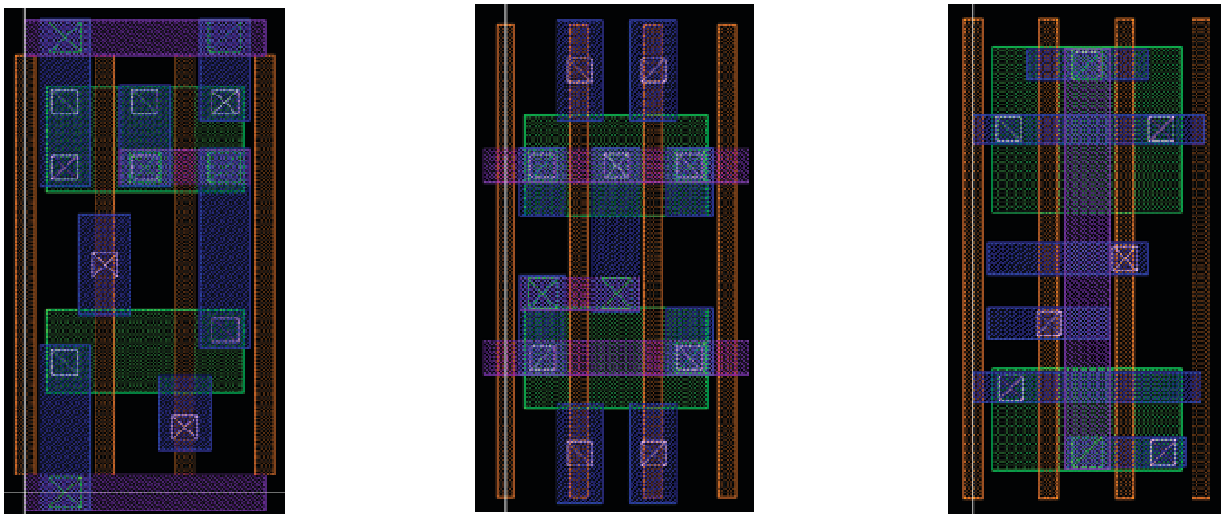



**Figure 3: Number-of-good dies are compared for 6T-β=1 and 6T-β=2 topologies at two different device lengths**

## 4. COST PER GOOD DIE ANALYSIS AT THE 32NM TECHNOLOGY NODE

First, the cost-per-good-die is used to analytically determine the optimum design style for a 32nm application specific integrated circuit (ASIC). This experiment was designed to evaluate the feasibility of applying regular design fabrics to ASICs at the 32nm technology node. We implemented a processor core using three different regular design fabric [12, 13, 15, 16] based on template libraries [15] created in 65nm technology and then scaled down to 32nm technology. We scale the layouts such that the dense metal pitches (M1 - M5) are 120nm, as is consistent with the pitches used for developing a simplified process in an earlier publication [13]. The description of the three different libraries is provided below:

- **7T pdBRIX template with M1 parallel to poly**
- **8T pdBRIX template with M1 parallel to poly**
- **8T pdBRIX template with M1 perpendicular to poly**



(a) (b) (c)

**Figure 4: (a) 7T pdBRIX template with M1 parallel to poly (b) 8T pdBRIX template with M1 parallel to poly (c) 8T pdBRIX template with M1 perpendicular to poly**

Illustrations for different design styles for templates on regular design fabrics is provided in Figure 4. For detailed discussion on regular design fabrics readers are encouraged to refer [12, 13, 15, 16]. All three designs used the same netlist post-synthesis and used a Cadence physical synthesis flow to find the minimum die size for the processor. The processor used for this exercise is optimized for density with very loose performance constraints. To make this analysis more practical it is necessary that we consider chip sizes close to that used in a typical product. We assume that the chip is a multiprocessor chip with 1024 processor cores arranged in a 32-by-32 array. The yield and effective areas for the three different implementations are shown in Table I. Due to the unavailability of a complete commercial standard cell library for such an analysis has led us to include estimates for a library using conventional standard cells (8T Irregular) and another library using regular poly and active (8T Restricted Design Rules) [21-23]. Both of these estimates are based on a small set of sample standard cells that are designed in-house. We make conservative assumptions that the 8T Irregular can achieve area parity with the densest template library, whereas a 8T Restricted Design Rule library will suffer a 10% area penalty. Although the 8T Restricted Design Rule library demonstrates improvement in yield compared to conventional standard cell libraries they do not follow the stringent pattern control achieved by templates on a regular design fabrics or SRAM and hence cannot guarantee 100% systematic yield. As a result we speculate that the systematic yield for 8T Irregular and 8T Restricted Design Rule to be 85% and 95% respectively. In contrast, the random defect limited yield is estimated more precisely using based on critical area extracted using YRS [32] and defect size distributions modeled for a process in early volume manufacturing. For the template libraries the final placed and routed GDS was used for the critical area extractions. Since we lack commercial standard cell library we estimate the critical area from a random placement of a small set of standard cells that are designed in-house for both the Irregular and Restricted Design Rule implementation.

| | Dimensions (mm) | Die Area (sq. mm) | Die Yield | Good Dies Per Wafer |
|---|---|---|---|---|
| **7T pdBRIX template: M1 parallel to poly** | **4.77 x 5.92** | **28.24** | **68.65%** | **1655** |
| **8T pdBRIX template: M1 parallel to poly** | **4.8 x 5.73** | **27.50** | **68.40%** | **1695** |
| **8T pdBRIX template: M1 perpendicular to poly** | **5.0 x 5.41** | **27.05** | **74.01%** | **1866** |
| **8T Irregular*** | **5.0 x 5.41** | **27.05** | **56.91%** | **1435** |
| **8T Restricted Design Rules*** | **5.25 x 5.68** | **29.82** | **64.96%** | **1482** |

**Table 1: Number-of-good-dies for design implemented using the three different libraries**

Based on the assumptions stated, the 8T pdBRIX template with M1 perpendicular to poly library shows significant improvement in the number of good-dies-per-wafer (Table I). We consider this as the optimal template library for such high density design and use it for further comparison of the cost-per-good-die

| | pdBRIX template | Irregular | Restricted Design Rule |
|---|---|---|---|
| Poly | Simple DPT | DPT | DPT |
| Active | SE w/1.2 NA | SE w/1.35 NA | SE w/1.2 NA |
| Contact | SE w/1.3 NA | DPT | DPT |
| Metal 1 | SE w/ 1.2 NA | DPT | SE w/1.35 NA |
| Via 1 | SE w/ 1.3 NA | DPT | SE w/1.35 NA |
| Metal 2 | SE w/ 1.2 NA | DPT | SE w/1.35 NA |

**Table 2: Summary of lithography choices for the critical (hard) layers for a 32nm technology node process**

To determine the cost-per-good-die, it is necessary that we first define the assumptions for lithography strategies for each one of these designs. The detailed discussion on the lithography system requirements for regular design fabrics as well as conventional standard cells at the 32nm technology node has been provided in an earlier publication [13]. In this paper we choose only to summarize lithography requirements for the different design styles in Table 2. During early volume production of the 32nm technology node the design requirements of the 8T Irregular library would require a widespread use of DPT. In comparison the use of pdBRIX templates or Restricted Design Rules can be successfully manufactured with 193nm single exposure (SE) immersion lithography for all critical layers. The Poly layer is an exception as we would require DPT to define the dense line-end configurations required for SRAM bitcells. By controlling the location and configurations of these line-ends in the pdBRIX templates we can employ a simplified DPT solution [12].

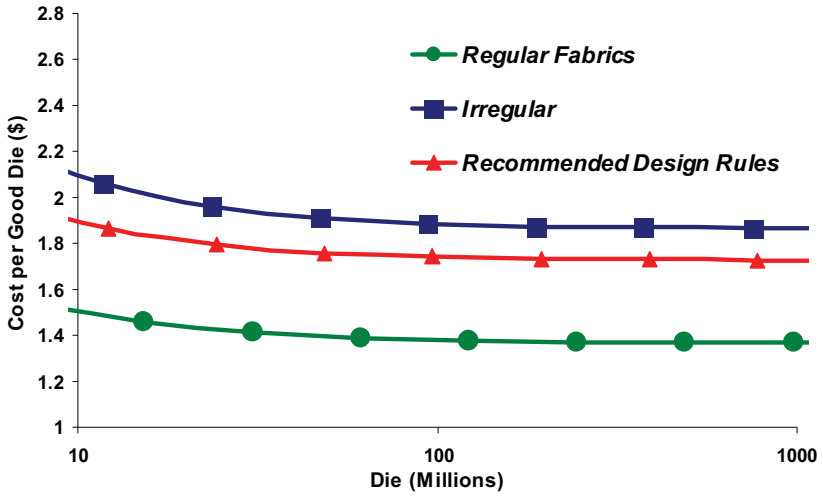



**Figure 5: Cost-per-good-die as a function of design style and volume at the 32nm technology node**

A total of 39 lithography steps are required for this process. Poly, Active, Contacts, M1-5 and V1-4 constitute the 12 critical (hard) layers. The remaining 27 non-critical layers are use single exposure lithography. 5 of these are use 193nm dry systems (med) and the remaining 22 use 248nm lithography system. The cost-per-good-die is plotted in Figure 5 using the nonlitho cost-per-wafer of $2000. The Restricted Design Rule library enabled a 10-20% cost-per-good-die reduction compared to the Irregular layout style across various product volumes. The use of pdBRIX templates library enables an addition 26% to 30% cost-per-good-die reduction. The results demonstrate the apparent advantage of using a more regular standard cell library compared to conventional standard cells in spite of the area increase. Most of all a strong case for the use of template based design using regular design fabrics has been made at the 32nm technology node. This result can also be stated in real dollar terms. If this product requires 100 million dies the use of the optimized

pdBRIX template library can enable a cost saving of $0.368 per die as compared to best standard cell implementation. This is a savings of $36.8 million over the lifetime of the product or an effective savings of $687 per wafer. For product volumes of 1 million units the savings are even more astronomical. A cost saving of $0.61 per die or $610,000 over the lifetime of the product is observed. This is an effective savings of $1138 per wafer.

## 5. DEVELOPMENT OF THE 22NM TECHNOLOGY NODE

We now discuss the application of the cost-per-good-die for technology development at the 22nm technology node. The 22nm node presents a challenge for conventional scaling practices. Although significant improvements have been made in all facets of EUV lithography in recent years, it will not be ready for volume production of the 22nm technology node [24, 30]. In effect lithographers will have to devise processes that extend the application of 193nm immersion tools for one more generation. DPT will have to be widely employed and it should be no surprise that lithography equipment vendors are successfully mobilizing resources to meet the 3nm misalignment target required for DPT at the 22nm technology node [27]. However, the increasing cost from DPT is yet unsettling to most IC manufacturers. In this section we evaluate other alternate lithography solutions to determine a process solution that would minimize the cost-per-good-die. IBM Microelectronics and Mentor Graphics are jointly developing the source mask optimization (SMO) technology to reduce the need for DPT [24]. Although these techniques have been widely publicized as an alternate to the high costs of DPT the costs of such "computational scaling" techniques has never been published. In this paper we would be the first to estimate these costs in our integral model for the cost-per-good-die. Others has been proposing the use of massively parallel e-beam (MEBM) to minimize the lithography costs for low to medium volume products [28]. Unfortunately, these MEBM techniques are already having trouble meeting the aggressive 30 wph throughputs initially quoted [28, 33]. Other researchers have also been proposing the use of simplified double patterning by using grating based lithography solutions such as HOMA [25] and COOL [18]. The viable lithography solutions for critical layers at the 22nm technology node and associated assumptions are documented in Table 3 as well as discussed below:

• Standard double patterning (DPT): This is front runner of all technologies. Double patterning has been already used in production from the 45nm technology node [18] and it employs a double exposure double etch technique. We estimate the costs based on the data published by TSMC [28].

• Source mask optimization (SMO): As the alternative to the high cost DPT, SMO was first demonstrated by Rosenbluth et al [31]. In these published results SMO has been successfully applied to memory patterns to enable significant improvement in process window and hence achievable resolution in production IC processes. It has been argued that the use of SMO requires a limited set of patterns to ensure that the optimization routines can converge to solutions that provide improved process windows. In support of this claim it has been shown that sending all arbitrary patterns through a SMO optimization would result in the conventionally used annular source [35]. In other words having a very diverse pattern set will not enable any improvement in process window regardless of the quality of the optimization algorithms or extent of computing resources available. We assume that designs that would be used with the SMO technology would have limited pattern diversity similar to that of designs created using regular design fabrics. To assist with our analysis on SMO based technologies we make we make moderate predictions on both the cost of a new pixilated source and computing resources required.

• Grating based lithography (GRATINGs): Another solution to avoid the increasing costs of DPT is to enable a simplified double patterning by using a simple grating or array as the first patterning step and following that with a cheaper lithography solution [18, 25, 26]. This technique has two distinct advantages. It can use a resolution optimized first exposure step. Here we consider the low cost interference based lithography [25, 26] for the first patterning step. The second patterning step is only required to make small openings or deviations to the regular structure printed during the first patterning step and as a result can use a simplified lithography solution. We consider using an older generation optical lithography solution for the second patterning step. For 22nm we can even employ a dry 193nm lithography tool for the second patterning step. Once again there is a lack of published data on the processing and tool costs for full wafer interference lithography systems and we make moderate assumptions to facilitate our analysis. We assume that the integration of the interference tool to costs an extra $10 million to the cost of existing 193nm immersion lithography tool, the processing costs for the interference lithography step to be equivalent to that of single exposure immersion optical lithography in addition to the processing costs of the respective second patterning step, and that the throughput is limited by the second patterning step.

| | SMO | GRATINGs | Single Exposure (193i) | DPT |
|---|---|---|---|---|
| Tool ($ mill) | 50 | 40 | 50 | 75 |
| CPU usage (khrs) | 100 | 5 | 20 | 40 |
| Processing/layer ($) | 16 | 31 | 16 | 32 |
| Maintenance ($ mill) | 5 | 4 | 5 | 7.5 |
| Mask Cost ($ K) | 80 | 40 | 80 | 160 |
| Source Cost ($ K) | 200 | 0 | 0 | 0 |
| Throughput (wph) | 130 | 100 | 130 | 100 |

**Table 3: Summary of lithography cost assumptions for the 22nm technology node**

A comparison of the per layer lithography costs for the various lithography strategies is shown in Figure 6. On the x-axis we plot the average wafer volume per product of the technology node. We notice that the costs per lithography layer flatten out beyond the 20,000 wafer target and hence choose to limit the x-axis to 20,000 wafers. The cost of a single exposure 193nm immersion process (Single Exp Optical) is plotted as the reference. We conclude that even at high volumes one can expect a 150% increase in patterning costs per critical layer by employing DPT. If the aggressive targets of interference based lithography are met it presents the only solution that can match the per layer patterning cost of 193nm immersion single exposure systems across all product volumes. At present SMO is the most practical solution for Fabs that have a large percentage of medium and high volume products.

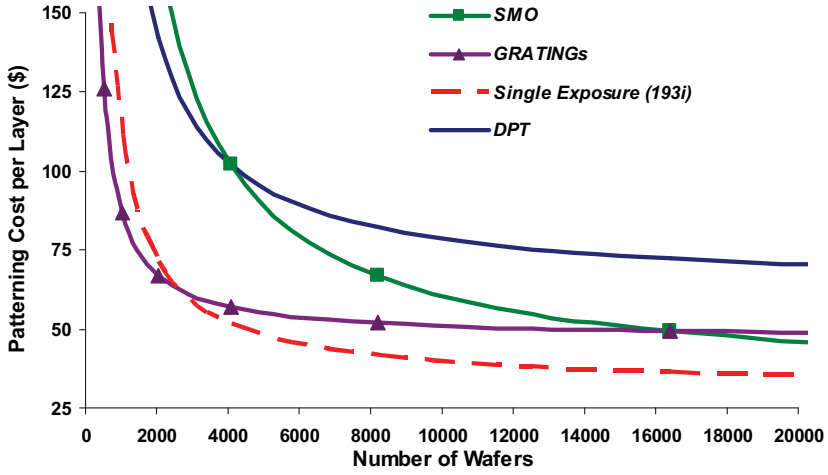



**Figure 6: Patterning cost per layer for the 22nm technology node**

We take the next incremental step in the analysis and calculate the total cost-per-wafer for a process technology with 39 lithography steps. We consider a combination of lithography choices for the 12 critical layers, i.e. Poly, Active, Contacts, M1-M5 and V1-V4. The various combinations of critical lithography steps for these critical layers are summarized in the table V. The remaining 27 noncritical layers are exposed using single exposure lithography. 5 are exposed using 193nm immersion systems, 10 are processed using 193nm dry lithography and a 248nm system is sufficient for the remaining 12 layers. The total lithography-cost-per-wafer for the different lithography strategies is illustrated in Figure 7. As a reference we plot the total lithography cost for patterning all critical layer for the 32nm technology node process (32nm). This reference plot serves as an indicator of the increase in total lithography-cost-per-wafer as we move from the 32nm to the 22nm technology node. The use of DPT for all critical layers can increase the lithography costs by 28% at high wafer volumes. In comparison the introduction of GRATINGs based lithography for all critical layers leads to a much lower 11% increase in total lithography costs at high wafer volumes. The readers are once again cautioned that comparing such lithography expenses independent of their implications on design is shortsighted. The implications of

these lithography choices on design styles are now presented. It is noticeable that the introduction of design modifications required for interference based lithography are most severe on the Active and Contact layer. We hence consider solutions that selectively introduce standard DPT along with interference based lithography solutions to find solutions that will minimize the cost-per-good-die.

| Experiment | Poly | Active | Contact | Metal 1-5 | Via 1-4 |
|---|---|---|---|---|---|
| 32nm | DPT | SE | DPT | SE | SE |
| DPT only | DPT | DPT | DPT | DPT | DPT |
| SMO with Selective DPT | DPT | SMO | DPT | SMO | SMO |
| GRATINGs Only | GRATINGs | GRATINGs | GRATINGs | GRATINGs | GRATINGs |

**Table 4: Summary of experiment with different lithography choices for critical layers**

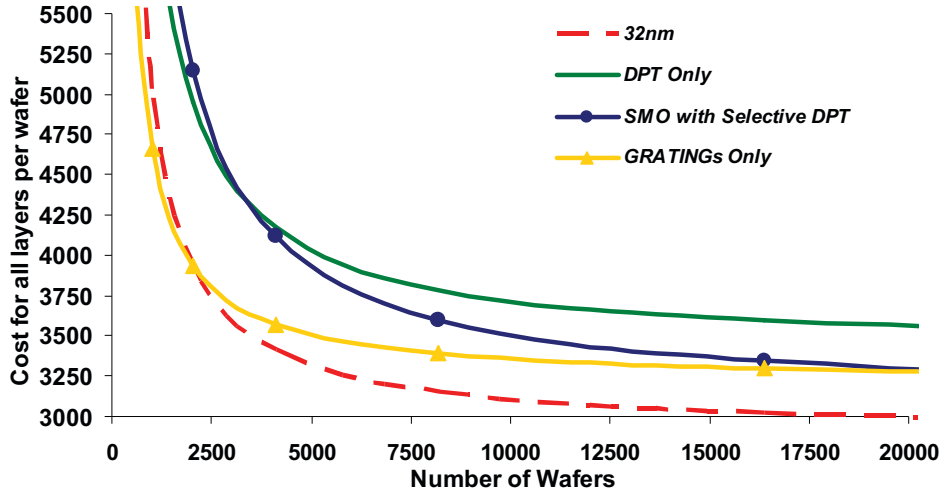



**Figure 7: Cost-per-wafer at the 22nm technology node**

The cost-per-good-die is computed for a 64MB SRAM chip. The array architecture consists of 32 columns by 16 row matrix of a 1Mb sub-arrays. Each of the 1Mb sub-array is designed to have sufficient redundancy to correct for all type of random failures except shorts between VDD and GND power rails. The area of the bitcell is computed by using a parameterized model and anticipated design rules for the different lithography strategy. A different parameterized model is created based on the unique design restrictions required for the different lithography choices. We assume a 50% periphery to bitcell ratio for the array. Yield is calculated by using a parameterized model that calculates the defect limited yield for a VDD and GND short for the specific design rules for each one the bitcells used, taking into account the redundancy offered by the array architecture. The die area and estimated yield is listed in Table 5.

| Experiment | Bitcell Area (sq. um) | Die Dimension (mm) | Die Yield | Good Dies per Wafer |
|---|---|---|---|---|
| 32nm | 0.1714 | 13.84 x 9.31 | 65.68% | 335 |
| DPT only | 0.0806 | 9.77 x 6.21 | 66.23% | 736 |
| SMO with Selective DPT | 0.0806 | 9.77 x 6.21 | 66.23% | 736 |
| GRATINGs Only | 0.1792 | 10.86 x 12.41 | 65.74% | 318 |
| GRATINGs with Contact DPT | 0.0896 | 10.86 x 6.20 | 64.47% | 643 |
| GRATINGs with Active & Contact DPT | 0.0806 | 9.77 x 6.21 | 66.23% | 736 |

**Table 5: Summary of experiment with different lithography choices at 22nm technology node**

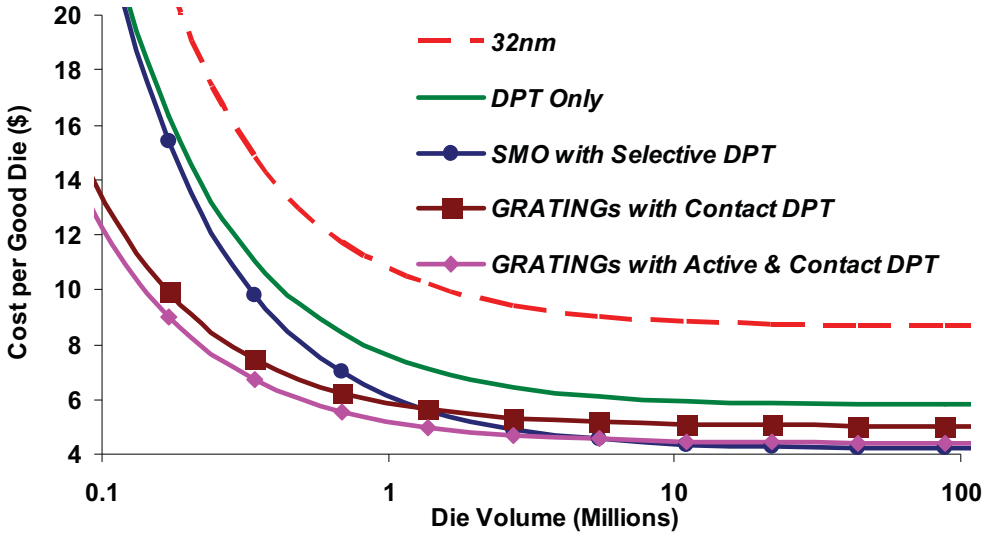



**Figure 8: Cost-per-good-die vs product volume for different lithography choices for critical layers at the 22nm technology node**

The chosen design rules predict the reference 32nm bitcell to have an area of 0.1714 sq. um. The use of DPT at 22nm technology node for all critical layers allows the use of the same bitcell topology as was used for 32nm with an effective area of 0.0806 sq. microns. Using GRATINGs for all critical layers requires increasing the poly pitch (height) by 100% to enable contacts on a manhattan grid as well as increasing the width of the bitcell by 11% to allow for diffusions (active) to be on grid. This is effectively a 122% increase in the bitcell area. By introducing DPT for the Contact layer, this area penalty is reduced to 11%. Further the area penalty can be entirely negated by additionally introducing DPT for the Active layer. We also consider an SMO Optimized solution where Poly, Contact and M1 are patterned using DPT that is designed to meet the same bitcell area as the All DPT solution. Combining these different cost components, we determine the cost-per-good-die (Figure 8). At 22nm we predict the non-litho related wafer costs to be at $2200; a 10% increase in non-litho related wafer costs as compared to 32nm. Introducing DPT for the contact layers reduces the cost-per-good-die significantly but the cost per-good-die is only minimized when we use DPT for both contact and Active layers as shown in GRATINGs with Contact and Active DPT approach. Although this approach shows a lower cost-per-good-die compared to All DPT across all product volumes, it would only be the preferred solution for low volume products, whereas as an SMO Optimized solution would be preferred for medium to-high volume products. The cross over occurs at 10-20 million dies. At low volumes (1 Million dies) the GRATINGs based solution can enable a 60% cost reduction compared to all DPT. For Medium-to-high volume products (20 Million dies) a SMO based solution demonstrates a 38% cost reduction compared to all DPT. In other words all DPT will be cost competitive only when it can achieve 38% design density reduction compared to an SMO optimized solution. Our analysis shows that this crossover point appears at a design pitch of 62nm.

We further quantify the benefit from our analysis in dollar terms. At low volumes (1 million dies) the GRATINGs based solution demonstrates a $2.18 savings per die as compared to All DPT approach. This corresponds to a $1485 in savings per wafer at low volume. For a medium volume product (10 million dies) the GRATINGs based solution demonstrates a $1.57 savings per die that corresponds to a $1198 in savings per wafer, when compared with All DPT. Finally, at medium,-to-high volume products (20 million dies) the SMO Optimized solution is preferred, which demonstrates a savings of $1.59 that corresponds to a savings of $1205 per wafer as compared to the All DPT approach.

## 6. CONCLUSION

In the lack of cost effective techniques to technology scaling we present means of achieving economic technology scaling by process-design co-optimization. Using the cost-per-good-die metric we can analytically determine the

optimum process and design solutions that will extend economical scaling to the sub-32nm technology nodes. Using the cost-per-good-die we have determined the optimum design library for the 32nm technology node. With a pdBRIX template library we estimate a saving of $0.61 for low volume products and a savings of $0.368 for high volume products compared to the best standard cell library. We also demonstrated the selection of the optimal lithography strategy for the 22nm technology node by comparing the cost-per-good-die for a 32MB SRAM array. We employ a parameterized model to compute the area and random defect limited yield of the array and use it to evaluate the cost-per-good-die. We have shown that the use of the Intf Optimized approach is recommended for low to medium volume products and the use of the SMO Optimized approach is recommended for high volume products. Using the suggested lithography approaches a cost-per-good-die reduction of up to $2.18 for low volume and $1.59 for high volume products can be achieved.

## ACKNOWLEDGMENT

This work was supported in part by the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity. The assistance of Dr. Julian Uscherschon and Weifeng Zhao at PDF Solutions for running YRS and estimating critical areas is appreciated. The authors would also like to acknowledge all colleagues at Carnegie Mellon University who have directly and indirectly contributed to this work.

## REFERENCES

1. G. E. Moore, “Cramming More Components onto Integrated Circuits”, Electronics, Vol. 38, No. 8, 1965
2. "Excerpts from A Conversation with Gordon Moore: Moore’s Law", Intel Corporation (2005). Retrieved 2006-05-0
3. W. Maly, “IC Design in High-cost Nanometer-technologies Era”, Proc. 38th Design Automation Conference, pp 9-14, 2001
4. W. Maly, et al, “Design for Manufacturability in Submicron Domain”, Proc. 1996 IEEE/ACM International Conference on Computer-aided Design, pp. 690-697, San Jose, California, 1997
5. W. Maly, “Cost of Silicon Viewed from VLSI Design Perspective”, 31st Design Automation Conference, pp 135-142, 1994
6. W. Maly, "Prospects for WSI a Manufacturing Perspective," IEEE Computer Magazine, Vo1.25. No.4, pp. 356-392, April 1992
7. W. Maly and J. Deszczka, “Yield Estimation Model for VLSI Artwork Evaluation”, Electronic Letters, Vol. 19, No. 6, Mar 1983, pp. 226-227
8. C. H. Stapper, et al, “Integrated circuit yield statistics”, Proc. IEEE, vol. 71, no. 4, pp. 453-470, Apr.1983
9. T.Jhaveri et al, “Economic Assessment of Lithography Strategies for the 22nm Technology Node”, Photomask Technology 2009, Proc. SPIE, Vol. 7488, 2009.
10. J. Le, et al, “STAC: statistical timing analysis with correlation”, Proc. of 41th Design Automation Conference, pp 343–348, 2004
11. R. Menon et al, “Zone-Plate-Array-Lithography (ZPAL): Optical Maskless Lithography for Cost-Efficient Patterning”, Emerging Lithographic Technologies IX , Proc. SPIE, Vol. 5751, 2005
12. T. Jhaveri, et al. “OPC Simplification and Mask Cost Reduction Using Regular Design Fabrics”, Optical Microlithography XXII, Proc. SPIE, Vol. 7274, 2009
13. T. Jhaveri, et al, “Enabling Technology Scaling with In Production Lithography Processes”, Optical Lithography XXI, Proc. SPIE, Vol. 6924, 2008
14. PhD Thesis (2009), T. Jhaveri, Carnegie Mellon University, 2009
15. L. Liebmann, et al, “Simplify to survive: prescriptive layouts ensure profitable scaling to 32nm and beyond”, Process Design Integration, Design for Manufacturability through Design-Process Integration III, Proc. of SPIE, Vol. 7275, 2009
16. T. Jhaveri, et al, “Maximization of Layout Printability/ Manufacturability by Extreme Layout Regularity”, JM3 06 (03), 031011, (July– September 2007)
17. A. K. K. Wong, Resolution Enhancement Techniques in Optical Lithography, SPIE press, Bellingham WA, 2001
18. Y. Bordovsky, “Marching to the Beat of Moore’s Law”, Advances in Resist Technology & Processing, Proc. of SPIE Vol. 6153, 2006
19. M. LaPedus, “Intel drops 157-nm tools from lithography roadmap”, EETimes, 22 May 2003, http://www.eetimes.com/news/semi/showArticle.jhtml?articleID=10801799

20. D. Lammers, “Intel: 'EUV Facts Don't Add Up' for 22 nm in 2011”, Semiconductor International, 22 April 2008, http://www.semiconductor.net/article/CA6553758.html
21. L Liebmann, et al, “High-Performance Circuit Design for the RET enabled 65nm Technology node”, Design and Process Integration for Microelectronic Manufacturing II, Proc. of SPIE, Vol. 5379, 2004
22. M. Lavin, F.L. Heng, and G. Northrop, “Backend CAD Flows for Restrictive Design Rules”, Proc. of ICCAD-2004, pp 739–746, November 2004
23. L. Liebmann, “Layout Impact of Resolution Enhancement Techniques: Impediment or Opportunity?”, Presentation at 2003 International Symposium on Physical Design, http://www.ispd.cc/slides/ispd2003_slides/07_1_liebmann.pdf
24. “IBM Develops Computational Scaling Solution for Next Generation 22nm” Semiconductors, 17 September 2008, http://www.physorg.com/news140881068.html
25. M. Fritze, et al, “Extending 193nm Immersion with Hybrid Optical Maskless Lithography”, Solid State Technologies, Vol. 49, Issue 9, http://www.solidstate.com/display_article/272044/5/none/none/Feat/Extending-193nmimmersion-with-hybrid-optical-maskless-lithograph
26. B. Smith, “Alternative Optical Technologies: More than Curiosities?”,Optical Microlithography XXII, Proc. SPIE, Vol. 7274, 2009
27. Mircea Dusa, et al, “Pitch Doubling Through Dual – Patterning Lithography Challenges in Integration and Litho Budgets”, Optical Microlithography XX, Proc. SPIE, Vol. 6520, 2007
28. B. Lin, “Marching of the Lithography Horses: Electrons, Ions and Photons: Past, Present and Future”, Optical Microlithography XX, Proc. SPIE, Vol. 6520, 2007
29. W. Maly, Y. Lin, M. Marek-Sadowska, “OPC-Free and Minimally Irregular IC Design Style”, Proc. 44th Design Automation Conference, pp. 954-957, 2007
30. D. Lammers, “Intel: 'EUV Facts Don't Add Up' for 22 nm in 2011”, Semiconductor International, 22 April 2008, http://www.semiconductor.net/article/CA6553758.html
31. A. E. Rosenbluth et al, “Simultaneous Mask and Source Patterns to Print a Given Shape”, Optical Microlithography XIV, Chris Progler, Editor, Proc. SPIE 4346 (2001)
32. YRS Software, www.pdf.com
33. E. Seevinck, F.J. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," JSSC, Oct. 1987.
34. N. Giercyznski, B. Borot, N. Planes, and H. Brut, "A new combined methodology for write-margin extraction of advanced SRAM," IEEE Microelectronic Test Structures, Mar. 2007.
35. J. Wang, S. Yaldiz, X. Li, and L.T. Pileggi, "SRAM parametric failure analysis," DAC, July 2009.
36. M. Khellah, et. al. "PVT-variations and supply-noise tolerant 45nm dense cache arrays with Diffusion-Notch-Free (DNF) 6T SRAM cells and dynamic multi-Vcc circuits," VLSI Circuits, June 2008.