Are Defect-Tolerant Circuits with Redundancy Really Cost-Effective? Complete and Realistic Cost Model

Yves Gagnon
Département de Génie Physique
École Polytechnique
Montréal, Québec, Canada, H3C 3A7
gagnony@grm94.polymtl.ca

Yvon Savaria
Département de Génie Électrique
École Polytechnique
Montréal, Québec, Canada, H3C 3A7
savaria@vlsi.polymtl.ca

Michel Meunier
Département de Génie Physique
École Polytechnique
Montréal, Québec, Canada, H3C 3A7
meunier@phys.polymtl.ca

Claude Thibeault
Département de Génie Électrique
École de Technologie Supérieure
Montréal, Québec, Canada, H3C 1K3
thibeault@ele.etsmtl.ca

Abstract

Yield enhancement for fault tolerant circuits with redundancy has been widely studied during the last years. Recent manufacturing technologies have brought out steady and significant improvement regarding contamination and defect density and have forced us to re-evaluate the economical advantages of circuits with redundancy. The main goal of this paper is to propose a realistic cost model for fault tolerant chips that includes manufacturing, test and reconfiguration processing steps. We demonstrate, with our model, how optimum and cost-effective redundancy levels can be determined, for a class of fault tolerant architectures, and we compare our results to the usual figure of merit calculation. We demonstrate that wafer test costs can have significant impact on the cost effectiveness of redundant implementation. This model also leads us to show that, for a optimized fault tolerant chip, the silicon cost tends to increase as a quasi-linear function of chip area. We finally demonstrate that, considering future fault densities expected for the next decade, economical advantages of redundancy will probably vanish for most integrated circuits when they are implemented with mature processes.

I. Introduction.

During the last two decades, active research has brought out several approaches in order to achieve tolerance to faults caused by manufacturing defects in integrated circuits. In the rest of the paper, we will use the term fault in the restricted sense of a fault resulting from one or more manufacturing defects. One of the most popular approaches to fault tolerance consists of adding redundancy. Since redundant implementation brings about additional manufacturing costs, economic feasibility should be addressed.

The first step in the evaluation of a cost-effective redundant architecture is yield enhancement modelling. Several models are available to evaluate the optimum number of redundant modules to be added to a circuit. Since we are dealing here with large area ICs, it would not be realistic to use Poisson distribution, because it has been clearly demonstrated that it does not reflect very well the yield of large chips[1]. Several models based on different distributions, like Price distribution[2], have been proposed, but most recent models are based on the well known negative-binomial distribution. Depending of the cluster size, large area clustering model [3],

small area clustering model [4] or unified negative-binomial distribution model [5] could be used. In addition to theoretical yield distributions, modelling approaches based on chip layout and defect statistics, and using a Defect to Fault Mapper(DEFAM), has been developed [6].

Even though yield enhancement has a major influence on cost of fault tolerant chips, there are also other significant parameters to be considered for a complete cost model. A simple cost model including usual testing and packaging costs has already been proposed [7]. This model is based on a global profit function that has been used to evaluate the optimum level of redundancy. However, for fault tolerant chips manufacturing, additional diagnostic testing cost, laser restructuring cost (if applicable) and additional wafer testing cost have to be taken into account. The purpose of this paper is to propose a cost model including all those aspects related to fault tolerant chip manufacturing. Our objective is then to develop expressions for relative costs, comparing standard chips costs to fault tolerant chips costs. This model will be used here to evaluate the optimum conditions for which redundancy implementation is cost-effective, the impact of additional testing cost and the global cost-effectiveness of redundancy.

In the next section, we will introduce fundamental cost model assumptions. In the third section, we will develop mathematical equations for absolute and relative manufacturing costs. The fourth one will expose and analyse different results obtained with this model, and conclusions will follow in the last section.

II. Basic assuptions.

Yield model calculation.

It is of interest to note that all yield models mentioned in the previous section are fully compatible with our cost model and can be embedded in it without major difficulties. In this paper, we will use the negative-binomial model for large area clustering[3]. Our choice has been mainly motivated by the good compromise between simplicity and experimental data fit [1]. This model has also the advantage of being independent of layout considerations, which simplifies our analysis. Koren and Stapper[3] proposed the following relation to evaluate the yield of "m" identical base modules with "r" redundant modules

$$Y_{(m+r)\,modules} = \sum_{i=0}^{r} \binom{m+r}{i} \sum_{j=0}^{i} (-1)^j \binom{i}{j} \left(1 + \frac{(m+r-i+j)\,\sigma A_m}{\alpha}\right)^{-\alpha} \qquad (1)$$

Where σ, A_m and α are fault density, module area and clustering parameter respectively.

Chip architecture.

Let us consider an idealized chip with a global area "A" divided in two main parts (Figure 1)*. The first part has "m" identical modules occupying an area "kA" (k=fraction of total chip occupied by modules) and the second part contains the rest of the circuit. To make the first part fault tolerant, we added "r" redundant modules and some control circuitry, for reconfiguration (Figure 2). Control circuitry has, of course, been included in the second part, since it cannot generally tolerate any fault. If we define "c" as the control circuitry overhead, we can write the two following relations which will be used in the next sections.

$$A_m = kA/m \qquad A_{2ndpart} = (1-k)A + cA \qquad (2)\text{-}(3)$$

*Fig. 1 and 2 are not layout representations but just a graphical way to visualize both kinds of circuitry.

Since Eq. 1 can be used only for the first part of the chip, yield for the second part will have to be calculated separately. To simplify that calculation, we assumed that the two parts are statistically independent. The global yield for fault tolerant chips is then given by:

$$Y = Y_{1stpart} \cdot Y_{2ndpart} \tag{4}$$

Standard chips manufacturing steps.

Manufacturing steps for standard chips are illustrated in Figure 3. In each step, we find the related cost and the expression representing the number of dice coming out successfully. Parameters N_w, Y, Y_{pack} are the number of dice per wafer, the manufacturing yield and the yield after packaging, respectively. It follows that $(Y_{pack}YN_w)$ is the overall number of good chip per wafer.

Fault tolerant chips manufacturing steps.

Manufacturing steps for fault tolerant chips are illustrated in Figure 4. In this process, we find a diagnostic testing step to identify reconfigurable chips, a reconfiguration step to activate redundant modules by laser restructuring or software reconfiguration, and a second wafer testing step for validating the chip after reconfiguration. Because of the redundancy and area increase, Y and N_w will be different for fault tolerant chips manufacturing flow and standard chips manufacturing flow. The tilda symbol "~" has been introduced to express this difference. According to this, $\tilde{Y}$ is the manufacturing yield for fault tolerant circuits and $\tilde{N}_w$ is the number of fault tolerant dice per wafer. The symbol Y_o is the fraction of chips already working after the first wafer test and Y_r represents the laser restructuring yield. It also follows that $(\tilde{Y}\text{-}Y_o)$ represents the number of circuits that can be repaired and that need reconfiguration. This model assumes that components that are tested good after first wafer test need no further processing step but packaging. This is realistic and cost effective, even though some systems may be designed otherwise.

As one can see, the only costs that have not been taken into account are the Non-Recurring Expenses (NRE). We assume that those costs are equivalent for standard and fault tolerant chips, which is reasonable as long as the control circuitry complexity is simple compared to the rest of the circuit. Our model covers both laser restructuring and software reconfiguration. Both methods have been used and described in several papers. It also seems that a mixed approach, including laser restructuring and software reconfiguration, offers a good trade-off regarding reconfiguration flexibility, silicon area added, speed, and power consumption[8].

III.Yield and cost model.

According to the previous choice for yield distribution, the expression for yield of a standard chip described in Figure 1 will be: (5)

$$Y = G\,(1 + \sigma A/\alpha)^{-\alpha}$$

where G is the gross defect yield. We introduce a gross defect yield parameter to represent the effect of large defects that will kill chips, even if they have been made to be fault tolerant. As shown later, the number of dice per wafer "N_w" is an important parameter for cost evaluation. It is thus essential to use a realistic value for "N_w". Based on geometrical arguments, we can obtain the following expression:

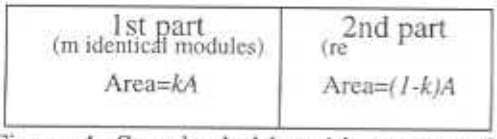

Figure 1: Standard chip with an area of A.

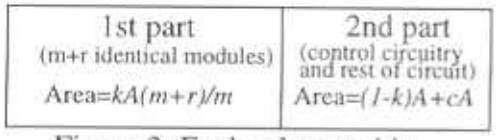

Figure 2: Fault tolerant chip.

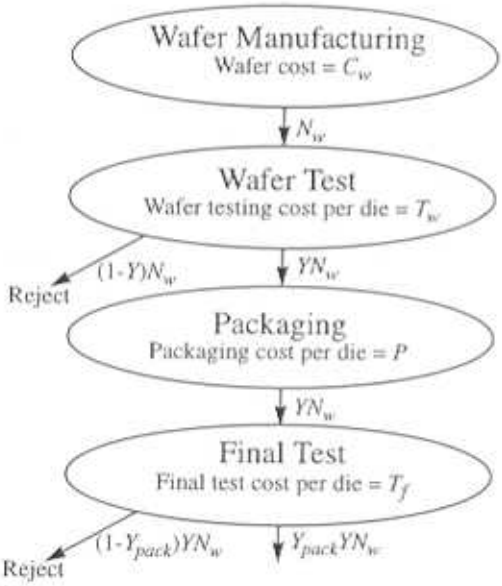

Figure 3:
Manufacturing steps for standard chips.

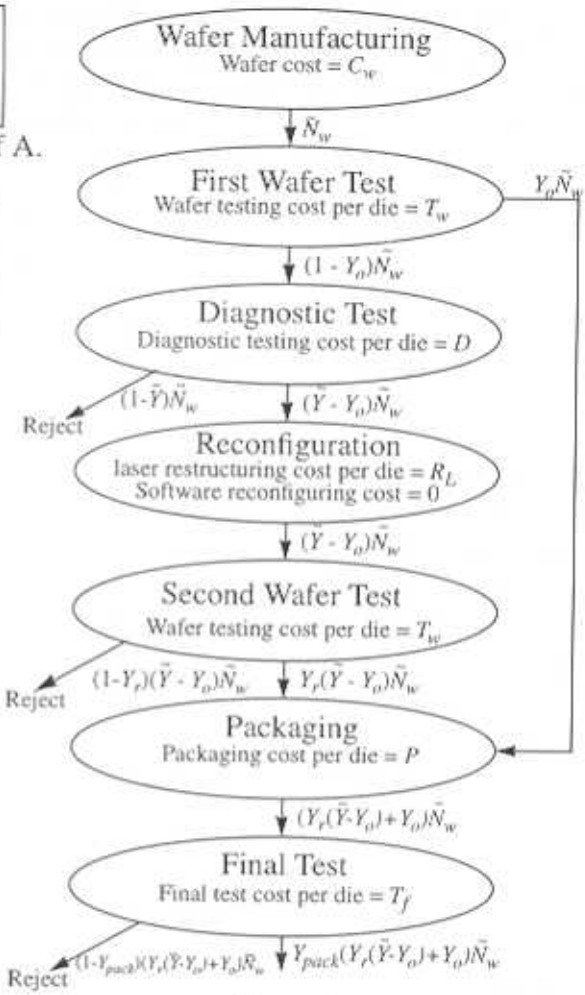

Figure 4:
Manufacturing steps for fault tolerant chips.

$$N_w = TRUNC\left[\left(1 - \frac{\sqrt{(2(L+S_L)(W+S_W)\pi)}}{2R}\right)\left(\frac{\gamma\pi R^2}{(L+S_L)(W+S_W)}\right)\right] \qquad (6)$$

where "L" and "W" are the dice length and width respectively, "S_L"and "S_W" are scribe channel widths and "R" is the wafer radius. The parameter "γ" represents the useful inner portion of a wafer. The left hand side parenthesis inside the brackets is a correction factor which models the area lost in the corners of peripheral dice. According to Figure. 3, we developed expressions for silicon cost "C_{si}", wafer test cost "$C_{wafertest}$", packaging cost "C_{pack}", final test cost "$C_{finaltest}$" and total cost "C" per good standard chip. These expressions are provided in appendix I.

As mentioned in section II, global yield for fault tolerant chips is calculated by Eq. 4. Yield for the first part will be obtained by substituting Eq. 2 in Eq. 1 and yield for the second part will be obtained by substituting Eq. 3 in a conventional negative-binomial distribution. Eq. 4 will then becomes:

$$\bar{Y} = G\left(\left(\sum_{i=0}^{r}\binom{m+r}{i}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\left(1+\frac{(m+r-i+j)\sigma kA}{\alpha m}\right)^{-\alpha}\right)\cdot\left(1+\frac{\sigma(1-k+c)A}{\alpha}\right)^{-\alpha}\right) \qquad (8)$$

Eq. 8 expresses the yield of fault tolerant chips as a function of the standard chip area. To find the expression for Y_o, we have to remember the meaning of Eq. 1. The second summation on the right hand side of Eq. 1 represents the probability to have a given combination of "i" faulty modules in "$m+r$" modules, and the first binomial coefficient is the number of combinations of "i" faulty modules in "$m+r$" modules. Multiplying these two expressions gives the probability to have any combination of "i" faulty modules on "$m+r$" modules. On the other hand, in an expression for Y_o, the binomial coefficient has to be the number of combination of "i" faulty modules among only "r" modules, because all "m" base modules have to be fault free. The expression for Y_o then reads:

$$Y_o = G\left(\left(\sum_{i=0}^{r}\binom{r}{i}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\left(1+\frac{(m+r-i+j)\,\sigma kA}{\alpha m}\right)^{-\alpha}\right)\cdot\left(1+\frac{\sigma(1-k+c)A}{\alpha}\right)^{-\alpha}\right) \qquad (9)$$

To evaluate $\tilde{N}_w$, we need new dimensions $\tilde{L}$ and $\tilde{W}$ of fault tolerant chips. From Figure 2, we find that the area of fault tolerant chips is $\tilde{A}=(kr+m+cm)A/m$. We then evaluate $\tilde{L}$ and $\tilde{W}$ from this equation, and substitute the result in Eq. 6. According to Figure 4, we developed expressions for silicon cost "$\tilde{C}_{si}$", wafer test cost "$\tilde{C}_{wafertest}$", reconfiguring cost "$\tilde{C}_{reconf}$" packaging cost "$\tilde{C}_{pack}$", final test cost "$\tilde{C}_{finaltest}$" and global cost "$\tilde{C}$" per good fault tolerant chip. These expressions are provided in appendix II. It follows that relative costs for fault tolerant chips are:

$$\Re_{si} = \frac{\tilde{C}_{si}}{C_{si}} = \frac{YN_w}{\left(Y_r\left(\tilde{Y}-Y_o\right)+Y_o\right)\tilde{N}_w} \qquad (11a)$$

$$\Re_{wafertest} = \frac{\tilde{C}_{wafertest}}{C_{wafertest}} = \frac{\left(\left(1+\tilde{Y}-Y_o\right)+(1-Y_o)\,(D/T_w)\right)Y}{\left(Y_r\left(\tilde{Y}-Y_o\right)+Y_o\right)} \qquad (11b)$$

$$\Re_{pack} = \frac{\tilde{C}_{pack}}{C_{pack}} = 1 \qquad \Re_{finaltest} = \frac{\tilde{C}_{finaltest}}{C_{finaltest}} = 1 \qquad \Re = \frac{\tilde{C}}{C} \qquad (11c,d,e)$$

To our knowledge, there exist no reliable function predicting growth of wafer test cost (T_w) and diagnostic test cost (D) with area. In absence of better knowledge on testing cost, the weakest hypothesis that we will need is to assume that growth of wafer test cost (T_w) and diagnostic test cost (D) are the same such a way that the term (D/T_w) reduces to a constant parameter and Eq. 11b is not function of any absolute cost. This allows us to show graphically relative silicon cost and wafer test cost (Eq.11a,b) as a function of chip area, without referring to any theoretical function to model test cost growth with area [7], which are strongly dependent of types of integrated circuits and marketing issues. This model has been implemented as a MATLAB program and simulation results, in case of interest, are discussed in the following section.

IV. Numerical results analysis.

Let us first state the technological parameters that we used for all the following simulations.

$\sigma = 0.3\ cm^{-2}$	$R = 100\ cm$	$\gamma = 0.90$
$\alpha = 2.5$	$S_L = 1\ mm$	$Y_r = 0.98$
$G = 0.95$	$S_W = 0.3\ mm$	$Y_{pack} = 0.92$

Minimum chip area leading to a cost-effective fault tolerant chip.

First, we assume that the chip is completely fault tolerant ($k = 1$), control circuitry overhead is 5% ($c = 0.05$) and diagnostic test is at least twice the cost of a normal wafer test ($D/T_w = 2$). According to these assumptions, relative silicon and testing cost (Eq. 11a,b) have been obtained as a function of the standard chip area for different values of "m" (Figures 5 and 6). We chose the amount of redundancy that minimizes relative silicon cost for A=300mm^2. Before analysing Figure 5, we should state some maximum acceptable value for the relative silicon cost to consider redundancy as being worth the effort. For this discussion, we assume that the maximum relative silicon cost for a cost-effective redundant implementation is 75%. We could then conclude that the minimum chip area to consider fault tolerance, for this example, should be between 200mm^2 and 400mm^2, depending on the number of identical modules "m". This kind of results have, of course, already been obtained by other authors, but we reproduced it here only as a sanity check for our model. Relative testing cost (Figure 6) is of course larger than 1 in some cases, because there are more testing steps in fault tolerant circuits (Figure 4). However, it is of interest that testing cost are in fact smaller for large chips because they are spread over the number of good dice which can become very low for standard chips. Even though wafer testing cost are usually much smaller than silicon cost, it is not necessarily the case in very hard to test chips, and this increased testing cost should be taken into account.

At the other extreme of the practical cases we could encounter, we have also considered a 50% fault tolerant chip (k=0.5) and similar results are depicted in Figures 7 and 8. From these graphs, based on our criteria, we conclude that there is no economical advantage to introduce redundancy if the fault tolerant part is less than 50% of chip area.

Optimum conditions for cost-effective fault tolerant chip.

In several papers, the usual way to estimate optimum conditions for cost-effective fault tolerant chips is to optimize a figure of merit[9] defined as:

$$f_m = \frac{\bar{Y}}{Y} \times \frac{A}{\bar{A}} \equiv \frac{1}{\Re} \qquad (12)$$

which is a kind of inverted relative cost. In Figure 9, we compare Eq. 12 (old model) to our proposed relative cost calculation Eq. 11e (new model). In this example, we used T_w=2\$, D=4\$, P=4\$, T_f=4\$, C_w=3000\$ and m=32. We first notice that both models lead to the same optimum amount of redundancy. However, the new model predicts a smaller reduction of the global cost and shows again the impact of additional tests for fault tolerant chips manufacturing. Our model also confirms the well known result that the optimum number of spare modules is larger for large chips, because yield improvement possibilities are better.

Cost increase as a function of area for fault tolerant chip.

In the ideal situation where yield would not be a function of chip area, silicon cost would increase as a simple linear function of chip area. However, yield strongly decreases with chip area, which causes silicon cost to increase as a non-linear function of area. Figure 10 shows the relation between absolute silicon cost and chip area, calculated from Eq. 7a and 10a. For this calculation C_w=3000\$ and k=1. The non-linear evolution of silicon cost with area is apparent in Figure 10. Our results also demonstrate that for a fully fault tolerant chip with a large number of identical elements and optimum amount of redundancy, silicon cost tends to increase as a quasi-

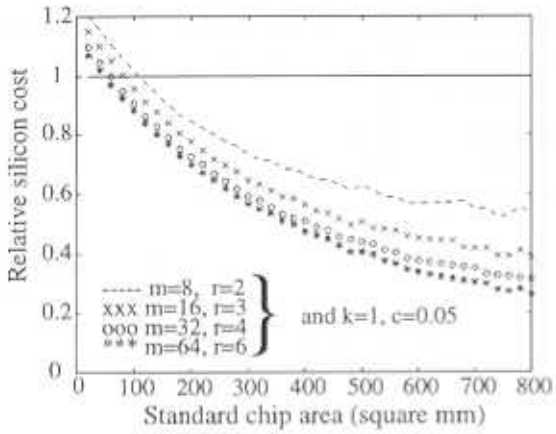

Figure 5:
Relative silicon cost for fault tolerant chips, k=1

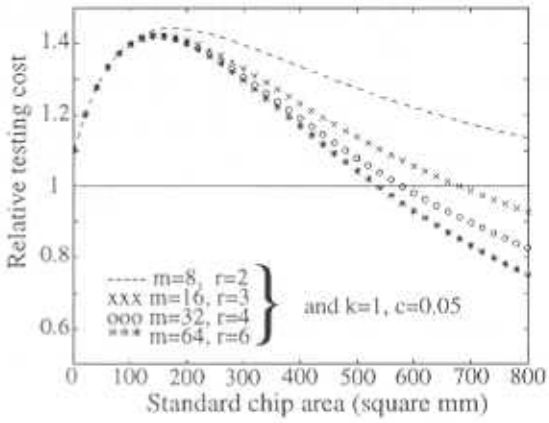

Figure 6:
Relative testing cost for fault tolerant chips, k=1

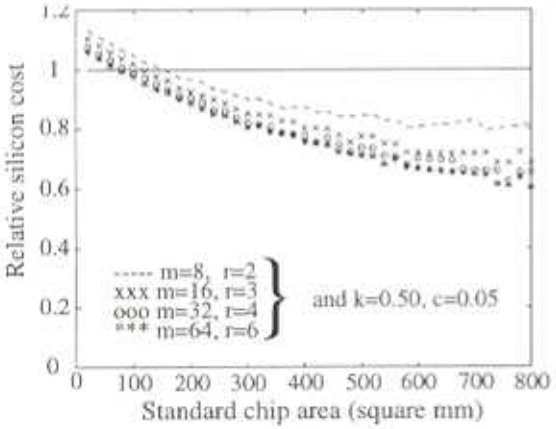

Figure 7:
Relative silicon cost for fault tolerant chips, k=0.5

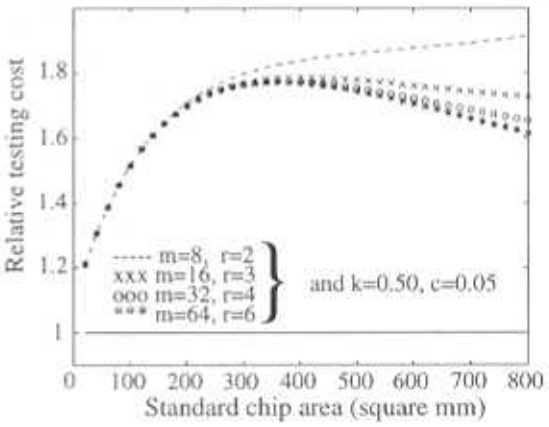

Figure 8:
Relative testing cost for fault tolerant chips, k=0.5

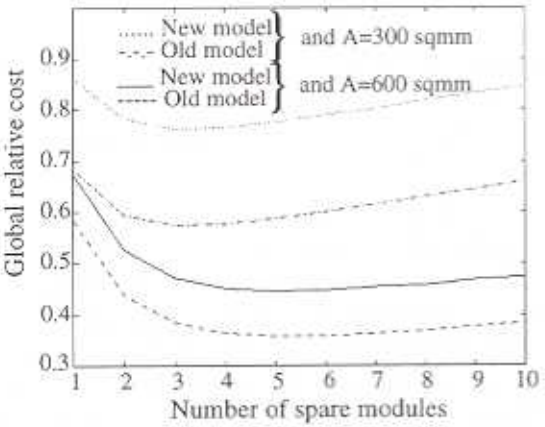

Figure 9:
Optimum number of spare modules

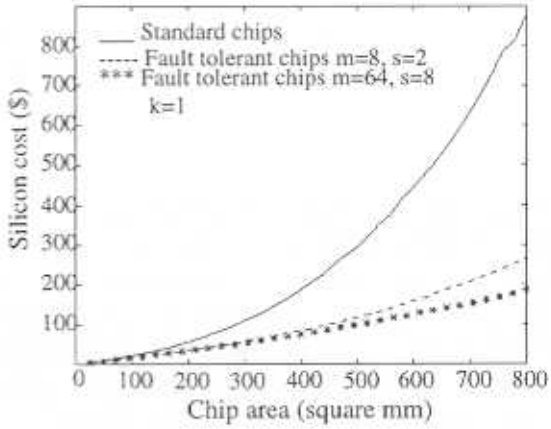

Figure 10:
Silicon cost for standard and fault tolerance chips

linear function of the chip area. It is of interest that this theoretical analysis regarding the difference between quasi-linear growth of silicon cost and the actual growth of a chip silicon cost without fault tolerance reflects the benefit of fault tolerance. The difference becomes significant only above a 200mm^2 area, which confirms the conclusion reached from the results in Figure 5. However, most circuits contain irregular logic, and even with optimal redundancy, a more than linear growth in silicon cost with area should be expected.

Future trends.

Future technologies based on in situ all dry processes and performed in cluster tools promise smaller fault density. This clearly impacts the potential benefit of fault tolerance. Figure 11 shows relative silicon cost for different fault density with m=64, r=8 and R=150mm (Eq. 11a). The National Technology Roadmap for Semiconductor from Semiconductor Industry Association (SIA)[10] is predicting that we should reach fault density as low as 0,01 cm^{-2} by 2007. Our purpose here is not to speculate on future fault densities but to illustrate that fault density strongly influences the minimum chip size for a cost- effective fault tolerant implementation. If fault densities reach values as low as 0,01 cm^{-2}, fault tolerance implementation according to the model of Figure. 2 will not be applicable any more even for very large integrated circuits. From Figure 11, we then conclude that, even with the optimistic result regarding the linear growth of silicon cost from the last section, cost-effective fault tolerance has a very uncertain future.

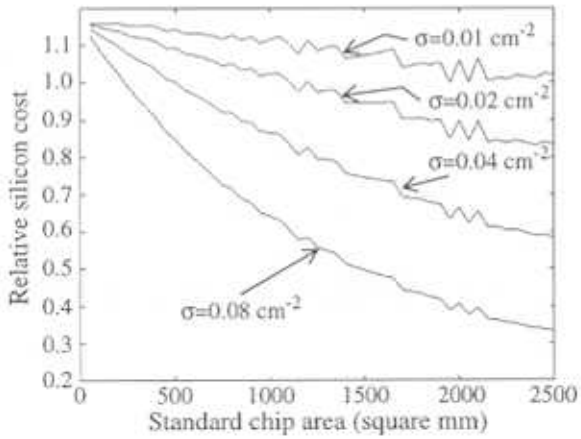

Figure 11:
Relative silicon cost for future defect densities

V. Conclusion.

In this paper, we proposed a realistic cost model for fault tolerant circuits with redundancy. The model allowed us to carry out relative cost for fault tolerant chips which can be used to evaluate optimum conditions for a cost-effective implementation. We showed that additional test costs, needed in the fault tolerant chip manufacturing, can have significant impact on fault tolerance benefits. We also showed that fully fault tolerant chips (as depicted in Figure 2) can lead to a quasi-linear silicon cost increase with chip area, but fault density improvements will probably prevent fault-tolerance from providing a significant manufacturing costs reduction, even for very large chips. Implementation of this model on computer with accurate yield calculation could become a useful tool for the evaluation of fault tolerance feasibility.

References.

[1] J. A. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing", *IEEE Trans. Semiconductor Manufacturing*, vol. 3, no. 2, pages 60-71, May 1990.
[2] W. Maly, "Feasibility of Large Area Integrated Circuit", *Wafer Scale Integration*, Ed. Earl E. Swartzlander, Chap. 2, pages 31-56, 1989.
[3] I. Koren, and C. H. Stapper, "Yield models for defect tolerant VLSI circuits", *Defect and Fault Tolerance in VLSI Systems*, vol. 1, I. Koren, Ed. New-York: Plenum, 1989, pages 1-21.
[4] C. H. Stapper, "Small-area fault clusters and fault-tolerance in VLSI circuits", *IBM J. Res. Develop.*, vol. 33, pages 174-177, March 1989.
[5] I. Koren, Z. Koren and C. H. Stapper, "A unified negative-binomial distribution for yield analysis of defect-tolerant circuit", *IEEE Trans. on Computers*, vol. 42 (6), pages 724-733, June 1993.
[6] D. D. Gaitonde, D. M. H. Walker and W. Maly, "Accurate Yield Estimation of Circuits with Redundancy", *International Workshop on Defect and Fault Tolerance in VLSI Systems*, pages 155-163, 1995.
[7] Z. Koren, I. Koren, "A Model for Enhanced Manufacturability of Defect Tolerant Integrated Circuits", *International Workshop on Defect and Fault Tolerance in VLSI Systems*, pages 81-92, 1991.
[8] G.H. Chapman, D.E. Bergen and K. Fang, "Wafer- Scale Integration Defect Avoidance Tradeoffs between Laser Links and Omega Network Switching", *International Workshop on Defect and Fault Tolerance in VLSI Systems*, pages 37-45, 1995.
[9] Mangir T., "Sources of Failures and Yield Improvement for VLSI and Restructurable Interconnects for RVLSI and WSI: Part I- Sources of Failures and Yield Improv. for VLSI", Proc IEEE, vol. 72, no. 6, pp. 690-708, June 1984.
[10] "The National Technology Roadmap for Semiconductor", Semiconductor Industry Association (SIA), "http://www.sematech.org/public/roadmap/doc/tbl_appb.gif" 1994.

Appendix I: Absolute costs for standard chips.

$$C_{si} = \frac{C_w}{Y_{pack} Y N_w} \qquad C_{wafertest} = \frac{N_w T_w}{Y_{pack} Y N_w} = \frac{T_w}{Y_{pack} Y} \tag{7a,b}$$

$$C_{pack} = \frac{Y N_w P}{Y_{pack} Y N_w} = P / Y_{pack} \qquad C_{finaltest} = \frac{Y N_w T_f}{Y_{pack} Y N_w} = T_f / Y_{pack} \tag{7c,d}$$

$$C = C_{si} + C_{wafertest} + C_{pack} + C_{finaltest} \tag{7e}$$

Appendix II: Absolute costs for fault tolerant chips.

$$\tilde{C}_{si} = \frac{C_w}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w} \tag{10a}$$

$$\tilde{C}_{wafertest} = \frac{\tilde{N}_w T_w + (1 - Y_o)\tilde{N}_w D + \left(\tilde{Y} - Y_o\right)\tilde{N}_w T_w}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w} = \frac{\left(1 + \tilde{Y} - Y_o\right)T_w + (1 - Y_o) D}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)} \tag{10b}$$

$$\tilde{C}_{pack} = \frac{\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w P}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w} = \frac{P}{Y_{pack}} \tag{10c}$$

$$\tilde{C}_{reconf} = \frac{\left(\tilde{Y} - Y_o\right)\tilde{N}_w R_L}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w} = \frac{\left(\tilde{Y} - Y_o\right) R_L}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)} \tag{10d}$$

$$\tilde{C}_{finaltest} = \frac{\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w T_f}{Y_{pack}\left(Y_r\left(\tilde{Y} - Y_o\right) + Y_o\right)\tilde{N}_w} = \frac{T_f}{Y_{pack}} \tag{10e}$$

$$\tilde{C} = \tilde{C}_{si} + \tilde{C}_{wafertest} + \tilde{C}_{reconf} + \tilde{C}_{pack} + \tilde{C}_{finaltest} \tag{10f}$$