INTERCONNECT AND CELL REDUNDANCY TRADEOFFS FOR WSI: AN FFT CASE STUDY

David Landis
Harry Nienhaus
Nitin Nigam
University of South Florida
Center for Microelectronics Research
Electrical Engineering Department
Tampa, FL 33620

Abstract

While redundancy is a prerequisite for defect tolerance in WSI, there is clearly a limit beyond which additional redundancy does not provide effective yield improvement. In addition, placement of redundancy also affects wafer wiring area overhead. WSI designs typically use a significant fraction of their silicon area for interconnections. The addition of redundancy for the purpose of reducing interconnect area also has the potential to increase overall system area efficiency. This paper presents preliminary results of a study which evaluates the effects of relative cell sizes and redundancy on overall wafer area. The effects of interconnect area, and the size and internal redundancy of the individual cells are evaluated. Finally, a case study is presented which applies these results to the design of a new 8-point FFT architecture which has been developed for wafer scale application.

Introduction and Background

The size of monolithic integrated circuits has generally been limited by the acceptable yield loss associated with defects within the manufacturing process. Advancing technology has reduced feature size and defect densities which allows higher levels of on-chip integration. However, the defect, yield, and cost relationships always place a physical limit on the maximum economical chip size. An alternative to traditional VLSI system assembly methods is to develop IC design and manufacturing technologies which are capable of tolerating defects. This is accomplished through the selective use of redundant components, along with a means to restructure each fabricated device to circumvent its unique defect pattern. One successful method for providing defect avoidance at the wafer level has been developed and demonstrated at the MIT Lincoln Laboratories [1], and transferred under DARPA support to the Center For Microelectronics Research (CMR) at the University of South Florida. This technique uses a laser to configure wafer level interconnections after wafer probe tests have identified defective components. Both additive and deletive interconnections can be made with this technology through the use of laser diffused link structures on the wafer. Alternative methods for restructuring of WSI designs are being investigated at CMR, including focused ion beam and laser enhanced deposition technologies.

* This research is being supported by the Defense Advanced Research Projects Agency under DARPA Grant No. MDA 972-98-J-1006.
While redundancy is a prerequisite for defect tolerance in WSI, there is clearly a limit beyond which placing additional redundancy does not provide effective yield improvement. A useful figure of merit for assessing the value of redundant circuitry is the yield to area ratio. Increasing redundancy will eventually result in a decrease of this figure of merit because yield improvement diminishes as cell area increases beyond some critical value. The optimum percentage of redundancy will vary according to the area of the individual cells in a WSI design. Another consideration in determining the amount and placement of redundancy concerns the wafer area required for cell to cell interconnections. It is not uncommon for a WSI design to utilize 50% or more of the available silicon area for interconnections. Thus, the addition of redundant cells for the purpose of reducing interconnect area also has the potential to increase overall system area efficiency.

**Cell Size and Placement for Minimum Wafer Area**

In order to assess the area efficiency of ULSI or WSI designs, a quantitative method is needed for evaluating the total wafer area in terms of cell and interconnect area. Total area will vary with the number, placement, and connections of cell I/O, as well as the cell sizes themselves. The following formulation ignores the functional aspects of the underlying system architecture and focuses directly on the cell size and interconnect requirements. This analysis considers a design scenario which involves only two distinct functional cell types. However, the general analysis framework could easily be extended for designs of more than two cell types.

Consider a system comprised of \( n \) cells (chips) of type \( C1 \) and \( n \) cells of type \( C2 \). Then the total wafer area can be defined as:

\[
A_{\text{wafer}} = (n \cdot A_{C1}) + (n \cdot A_{C2}) + A_w
\]

where \( A_{C1} \) and \( A_{C2} \) represent the areas of cells \( C1 \) and \( C2 \) respectively, and \( A_w \) is the total wiring interconnect area. Figure 1 illustrates the a generic layout of a wafer which contains two cell types. Let \( W \) represent the total number of wires in the buses which connect type \( C1 \) cells to type \( C2 \) cells, with \( P \) as the wiring pitch. In the example of figure 1 there are \( (1 \times 5) \) arrays of cells of type \( C1 \) and \( C2 \). Assuming that the cells are square allows the maximum length of the interconnect wiring to be calculated based on the maximum distance required to span the array:

\[
L_{\text{max}} = n \cdot \sqrt{A_{C1}} + (n-1) \cdot W \cdot P
\]

Next, consider only the region of wiring area between the array of type \( C1 \) cells and the array of type \( C2 \) cells. Each of the wires occupies an area \((W \cdot P)\). If we hold the size of cell \( C1 \) fixed, and allow the size of cell \( C2 \) to shrink, an increasing number of wires will be required in this wiring channel to accommodate the staggering of signal
lines. If $r$ is the ratio of the cell areas $A_{Cl}/A_{C2}$, then the minimum wiring channel width would be $(W P)/r$. This equation remains valid until the area ratio $r$ drops below $1/n$. Thus the total area occupied by this wiring region is:

$$A_w = L_{max} \cdot (channel~width) = \left[\frac{n \cdot \sqrt{A_{Cl}} + (n-1) \cdot W \cdot P}{W \cdot P} \right] / r$$  \hspace{1cm} [3]

Substituting this expression into equation [1] gives the total wafer area as the combined cell and interconnect wiring area.

A primary objective of this study is to evaluate the effects of relative cell sizes and redundancy on overall wafer area. Within the individual cell arrays, wiring channel requirements will not change with cell size or redundancy. Thus, we can group the interconnect area within the $C_1$ or $C_2$ cell arrays as part of the cell area without loss of generality. If $Y_i$ is the yield of cell $C_i$, then in order to get $n$ working cells we must place a minimum of $(n/Y_i)$ type $C_i$ cells on the wafer. Cell yield is predicted using the gamma (negative binomial) distribution, as described in [2]. Hence, the minimum total wafer area required to construct a functional system will be:

$$A_{wafer} = (n \cdot A_{C1}/Y_i) + (n \cdot A_{C2}/Y_2) + \left[\frac{n \cdot \sqrt{A_{Cl}} + (n-1) \cdot W \cdot P}{W \cdot P} \right] / r$$  \hspace{1cm} [4]

The actual area of cell $C_1$ (including redundancy and intra-array wiring) is $A_{C1}$. Let the non-redundant area required to implement the basic cell function be defined as $a_{01}$. The total cell area can then be divided into critical area and redundant area. The critical area is that part of the circuit where the presence of a defect constitutes a fault (the non-redundant hard core). The redundant area is that part of the circuit which has been replicated to provide defect/fault tolerance. Let the redundant area be $ra$, the critical area be $ca$, and the redundancy factor (ratio of total cell area to non-redundant area) be $r1$. Then, $ra = a_{01}/(r1-1)$; $ca = (2-r1)a_{01}$, and $A_{C1} = 2 \cdot ra + ca$. For the case of 10% redundancy a single fault can be tolerated in 20% of the cell area.

This analysis includes the implicit assumption that the cell is composed of separate individual modules which can be replicated to provide defect/fault tolerance within the cell. Next, consider a cell which is constructed from separate modules, each having redundant and non-redundant sections. The probability of obtaining a working cell is the probability of no defects in the entire cell, or of a defect in the non-critical area and none in the critical area. Let $Pr$ be the probability of having no defects on the entire cell, $Pr_{rr}$ be the probability of having a single defect in the non-critical area, and $Prc$ be the probability of having no defects in the critical area. Then the probability of getting a working cell can be estimated as $Yr = Pr + Pr_{rr} \cdot Pr_c$, where $Pr = \Gamma(x + s)^{(a^*d)/\bar{a})}/\Gamma(s^{(a^*d)/\bar{a}}) \cdot \Gamma(s)^{(1 + (a^*d)/\bar{a})}$, and likewise for $Pr_{rr}$ and $Prc$. Substituting into equation [4] produces the expression for total wafer area as:

$$A_{wafer} = \frac{nA_{C1}}{Pr_1 + Pr_{rr} \cdot Pr_{c1}} + \frac{nA_{C2}}{Pr_2 + Pr_{rr} \cdot Pr_{c2}} + \left[\frac{n \cdot \sqrt{A_{Cl}} + (n-1) \cdot W \cdot P}{W \cdot P} \right] / r$$  \hspace{1cm} [5]

where $Pr_1$, $Pr_{rr}$, $Prc_1$ are the yield values for cell $C_1$. If $r1$ is the redundancy for cell $C1$ and $r2$ is the redundancy for cell $C2$, then the value of $r$ in the previous equation can be substituted as $r = (r2-a_{02}) / (r1-a_{01})$. Figure 2 shows the behavior of equation...
plotting wafer area vs redundancy ($r_2$) for varying numbers of cell I/O ($W$) ranging from 16 to 32. As seen in the figure, the wafer area increases steadily with increasing redundancy for small values of $W$. However, as the interconnect requirements increase beyond 24, the area actually decreases with increasing redundancy up to a certain point. For the range of $W$ where area is seen to decrease, the interconnect area $A_c$ represents a significant fraction of wafer area, and the increase in redundancy reduces the required wiring channel area. Beyond the critical redundancy point, the area again increases because the total wafer area becomes dominated by the cell area.

Figure 3 shows an expanded scale plot of the effects of wiring area on the overall wafer area. In this figure, the array size is $n = 7$, $A_c$ is set at 0.5 and $A_c$ is initially set at 0.1. The redundancy of cell $C_2$ again varies from 1 to 2, representing a change in $A_c$ from 0.1 to 0.2. The minimum wafer area is clearly observed at the point where $r_2 = 1.42$.

**FFT Case Study**

An example of a WSI parallel pipelineable 8-point FFT implementation is reported in [5]. This architecture uses a single multiply-subtract-add cell to perform the required FFT computation. Four of these cells are needed to construct a two point butterfly, and 12 two point butterflies are then connected to construct the complete 8-point FFT. Figure 4 illustrates this conventional 8-point FFT configured in a 64-point pipeline. A total of 48 hardware multipliers are used for this architecture.
Our new 8-point FFT architecture, illustrated in figure 5, was derived from figure 4 using cut set graph theory to redistribute the multiplies [4]. It consists of a \((3 \times 4)\) array of multiplierless butterfly cells, with complex multiplications performed at the complex butterfly inputs by a separate complex multiplier cell. Seven of the eight inputs to the first butterfly stage require complex multiplications, as shown in figure 5. The internal multiplications are independent of pipeline length, position, and time. Three of these internal multiplications are by \(j\), which does not require a physical hardware multiplication. The other two have fixed twiddle factors and only two real multipliers are required to perform these complex multiplications. As a result, the total required number of hardware multipliers in this architecture is 32, a 33% reduction over the conventional FFT architecture.

Both global redundancy and local redundancy were considered in order to optimally utilize available wafer area. Cell yield estimates were calculated to determine the required levels of redundancy needed for successful restructuring of the wafer. The estimates in the following section are based upon an assumed defect density of 3 defects/cm\(^2\) using the negative binomial defect distribution [5], with a distribution shape parameter of 1.4 as recommended by [6].

As previously described, the two cell types required for this new FFT architecture are the complex multiplier and the complex multiplierless 2-point butterfly. The 2-point butterfly cell is relatively simple, containing only two 16-bit adders with associated scanable output registers. Designed under 2 micron MOSIS sCMOS ground rules, this cell will occupy 1.36\(\text{mm}^2\) resulting in an estimated cell yield of 96%. With such a high yield, only global redundancy is employed in the full wafer layout. Twenty-four 2-point butterfly cells are placed on the wafer for layout symmetry, however only 12 good cells are required in an operational 8-point FFT. The complex multiplier is a substantially more complicated cell, requiring four 16x16 bit parallel multipliers and two 24-bit adders. A straightforward implementation of this cell would occupy 32.6\(\text{mm}^2\). For this non-redundant implementation, cell yield is estimated to be 54%. Purely global redundancy for the complex multiplier would result in a very inefficient use of the wafer area because of the large number of spares required. It is easily observed that the majority of the active area required by the complex multiplier is consumed by the multipliers themselves. Thus, the occurrence of a single defect within this cell will likely render one of the multipliers faulty.

A fault tolerant complex multiplier cell has been designed to provide yield improvement through the incorporation of local redundancy. By adding a spare multiplier element, the cell can be configured to correctly perform the required complex multiplication if any four of the five multipliers are initially functional. Furthermore, if all multipliers are fault free, the spare element can be held in reserve.
to provide fault tolerance should a multiplier fail during system operation. The complex multiplier with redundancy will occupy 36.7 mm$^2$. The probability of obtaining a completely defect free cell with five functional multipliers is 61%. However, the true cell yield is the probability of obtaining no defects in the critical area and at least 4 good multipliers, which is estimated to be 60%. A test chip for the complex multiplier WSI cell has been designed for submission to MOSIS for fabrication. Figure 6 shows the layout of this test chip, with the five 16x16 multiplier cells clearly identifiable. The cell in the upper right hand corner of the layout is a Standard Test Interface (STI) which implements the IEEE 1149.1 test access port standard [7]. This interface is required to support our wafer level boundary scan design for testability methodology [8]. Twenty-four of these complex multiplier cells are placed on the wafer to insure that sufficient good cells will be available to construct an operational 8-point FFT. The overall wafer floorplan is illustrated in figure 7.

Figure 6 presents results of the yield model as applied to the complex multiplier cell of this FFT case study design. When the redundancy factor ($r$) equals 1, the entire cell area is critical in the sense that any defect will constitute a fault. However, as the redundancy increases the yield increases until at $r=2$ a single defect can always be tolerated, as seen in the $P_{rr}$ curve.
Looking at the overall yield curve, it is seen that as redundancy increases, the yield first increases, and after reaching a maximum at approximately $r=1.28$ it actually begins to decrease. This is expected because the increase in redundancy also results in a corresponding increase in overall cell area. For our complex multiplier cell design, the redundancy must be increased in units of $4.1 \text{mm}^2$ (the area of the 16X16 multiplier). Thus our predicted cell yield of 60% corresponds to the point on the curve where $r=1.125$, which is close to the true optimum yield value of 61.1% at $r=1.28$.

Summary

This extended paper summary has described the results of our investigation into the relationship between cell redundancy, yield, and interconnect area for WSI designs. A general analysis framework is presented which allows algorithmic computation of the optimum amount of redundancy from an area efficiency standpoint. The results of this theoretical study were applied to the case study of an 8-point FFT WSI design which uses a new architecture requiring only two functional cell types. The optimum design point and predicted cell yield are identified for this case study.

References:


