Yield Enhancement Considerations for a Single-Chip Multiprocessor System with Embedded DRAM

Markus Rudack Dirk Niggemeyer
Laboratory for Information Technology
Division Design & Test
University of Hannover
{*rudack, niggemeyer*}*@lfi.uni-hannover.de*

Abstract

A programmable single-chip multiprocessor system for video coding has been developed. The system is implemented in a high-performance 0.25 μm logic/embedded DRAM process. It integrates four processing elements, a total of 16 Mbit DRAM, and application specific interfaces. A hierarchical test strategy has been developed to test the different structures of the system such as processing elements and embedded DRAM. Logic testing is controlled by a fault tolerant BIST controller. The DRAM macrocells are supplied with integrated test facilities and word line redundancy, resulting in a yield of 99.0% for a 4 Mbit DRAM macro. To avoid soft failures, an SEC-DED error correction code (ECC) scheme for the DRAM has been realized. Even though the implementation of the ECC results in an area overhead of about 12%, the overall system yield is not decreased due to the effects of the ECC on defect tolerance of the memory. The 4 cm^2 multiprocessor system is suitable for utilization as a building block of a Large Area Integrated Circuit (LAIC).

1. Introduction

High-performance multimedia applications demand high processing power and high memory bandwidth, e.g., for real-time processing of high resolution video, as well as low power consumption. These applications benefit from the advances in Systems-On-Silicon which allow the realization of monolithic systems integrating digital circuitry with large memory arrays. Thus, restrictions of conventional systems in terms of bus widths, power consumption of pad drivers, or overall system size, can be overcome. The implementation and testing of such a system is described in this paper.

A special programmable video signal processing architecture has been developed and implemented that offers high flexibility and processing power to implement different video coding schemes, e.g., for video telephony [1]. For leading edge video coding, this architecture can be used as a processing node in a homogeneous multiprocessor system. To optimize the utilization of processing resources of this architecture and to demonstrate the capabilities of Systems-On-Silicon, a new monolithic system has been implemented integrating four processing nodes, embedded DRAM as high bandwidth frame memory, and application specific interfaces [2]. Consisting of more than 24 million transistors on a silicon area of 4 cm^2, this system is designed to be used as a building block for even larger multiprocessor systems realized as Large Area Integrated Circuits (LAIC). Thus, yield enhancement of certain elements is essential to manufacture the overall system with reasonable yield. Yield enhancement considerations and the hierarchical test strategy used for the system are the focus of this paper.

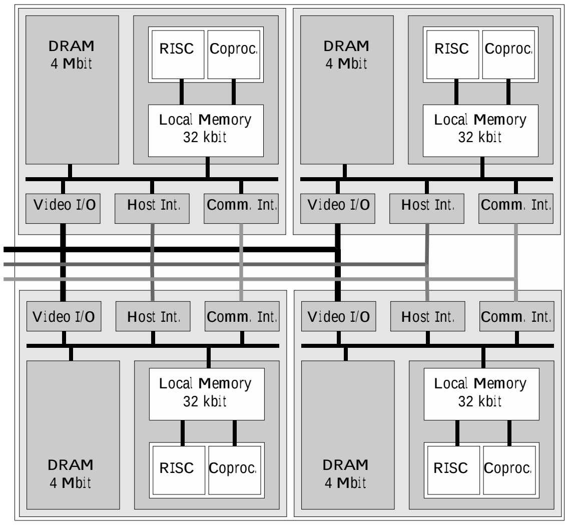

Figure 1. Single-Chip multiprocessor system consisting of four processing nodes

The remainder of the paper is organized as follows. In Section 2, the monolithic multiprocessor system is introduced and an overview of its architecture is given. The following sections deal with implementation issues of this system. Advanced standard cell logic and embedded DRAM show fundamentally different structures, though manufactured in the same process. A hierarchical test concept suitable for the different structures is implemented to provide the required fault coverage for realization of a Large Area Integrated Circuit. This is described in Section 3. The effects of applied redundancy schemes on the yield of the embedded DRAM and on the yield of the whole chip are discussed in Section 4. Integrating DRAM with standard cell logic introduces intermittent faults as a new fault mechanism to be considered by the system designer. To address these faults, an error correcting code (ECC) scheme is implemented with the DRAM under constraints mainly given by the available macros. This is also described in Section 4, as well as the effect of ECC on the yield of the chip. An outlook on the implementation of a Large Area Integrated Circuit is given in Section 5. Section 6 concludes the paper.

2. System Overview

The implemented system is designed to fulfill the computational requirements for video coding according to MPEG4 (simple profile @ CCIR601 resolution) [3]. The complete system consists of four processing nodes, each consisting of a processing element AxPe, its own 4 Mb frame buffer memory, and application specific interfaces. Communication between the processing nodes is possible via a dedicated interface (Figure 1).

The AxPe processing element is a coprocessor architecture adapted to hybrid video coding schemes. The RISC processor in each AxPe is used for computation of the medium level tasks and performs several control tasks. The coprocessor is used for fast processing of convolution-type low level tasks. Data transfer to the RISC processor and the coprocessor is carried out via a 32-kb local memory built of static RAM. The embedded DRAM is designed as a second level memory and will be used as frame memory and program memory. In this mode, no external memory is necessary, since three frames in CCIR601 resolution can be held on-chip with a total of 16 Mb DRAM.

The video interface of each processing node transfers video data between external devices and the embedded DRAM. It is initialized by software with a set of parameters which describe the region of the frame that is grabbed by the processing node. Thus, each processing node can process a different region of a larger frame independently most of the time. The host interface is integrated for transfer of coded video data and status information. Data exchange between the processing nodes is carried out by a communication interface with a distributed arbitration scheme.

3. Hierarchical test strategy

The implemented system consists of four identical processing nodes with embedded DRAM. The test concept is designed in a way that each processing node is tested independently and in parallel with the minimum number of test pins. Since this system will be the building block of an even larger monolithic system, a very high fault coverage of the system is essential for high reliability of a Large Area Integrated Circuit. This is achieved by a hierarchical and modular test strategy as described below.

The core of the test strategy applied for the AxPe processing element is a programmable defect and fault tolerant two-rail coded self-test controller that controls the test phases of the RISC processor and evaluates the test responses globally. Generally, in all modules of the RISC processor the system registers are replaced by self-testable BILBO registers [4], which can be employed as pseudo-random test pattern generators or as test signature analyzers.

The test signatures of the local self-tests are read out using multiple local scan chains resulting in a single test evaluation register per processing node. Instead of comparing the final signature to a golden signature, the following property of linear feedback shift registers is utilized. For each possible signature pattern stored in the register, a unique number of cycles is needed to get the inverse of the signature. XORing the original with the inverse results in all-1s for the correct signature. Therefore, it is only necessary to check for all-1s to create the pass/fail signal.

Thus, a flexible test evaluation procedure is implemented, that in case of a circuit extension, requires the adjustment of cycles needed for signature evaluation only. Likewise, in case of an insufficient fault coverage of the built-in self-test, the number of applied test patterns can be increased without circuit modifications. After finishing the RISC self-test, the modules of the low-level coprocessor and the local memory are tested sequentially by the RISC.

The embedded DRAM macrocells are equipped with a linear order march-like self-test feature that is utilized for manufacturing test. The test result is used to perform laser reconfiguration that exchanges faulty memory rows with spare rows. Since the dominant faults in single transistor DRAMs are proven to be intermittent soft errors, error correcting code circuitry has been added to the memory.

Defects in global interconnect systems are a major concern in large Systems-On-Silicon [5]. Therefore, the global interconnect systems are designed to be fault tolerant by providing spare bus lines. The global interconnect system is tested between the processing nodes by means of a fault tolerant scan path to supply the necessary information for reconfiguration. Reconfiguration is done by processing arrays of laser fuses and anti-fuses, replacing defective bus lines by spare bus lines.

To summarize, for each different structure such as logic, memory, and interconnects, suitable test approaches have been implemented that enable the hierarchical self-test of the individual processing node and the multiprocessor system. System performance is not influenced by insertion of these test measures, but rather it is determined by a trade-off between clock speed and required DRAM size.

4. Yield calculations

Yield is a major concern in designing systems of this complexity, since it is unlikely that defect-free chips as large as 4 cm^2 will be manufactured. Defects are even more likely in embedded DRAM processes, since these processes require more masks and more process steps than a standard logic process. Furthermore, the structures of standard logic and embedded DRAM are quite different, and so are the defect mechanisms and distributions of defects.

Redundancy is required to achieve working silicon, but different approaches are needed for logic and DRAM circuitry. The multiprocessor system introduced here employs a hierarchical redundancy scheme with spare rows for DRAM provided by the macrocells and spare buslines for the interconnect systems on the lower level. On the top level, complete processing nodes can be switched off if a defect occurs in the processor logic or if a defect in its DRAM cannot be repaired.

4.1. Yield of embedded DRAM

For the application specified above, 4.0 Mb DRAM configured as 128k×32 bit is required for each processing node. A 4-Mb macrocell is organized in 16×256 pages of 32 words of 32 bits in this case. An array of 256 pages is referred to as a book. The macrocells employ word line redundancy with four redundant pages for each book. Coding of failing addresses is done by fuses, which are processed after application of Built-In Self-Test for the memory array in the foundry. The yield Y_b of a book can be estimated by equations (1) to (3), which consider large area fault clustering [6].

$$Y_b = \sum_{k=M}^{N} P_N(k) \tag{1}$$

$$\text{with} \qquad P_N(k) = \binom{N}{k} \sum_{i=k}^{N} (-1)^{i-k} \binom{N-k}{i-k} y_i \tag{2}$$

$$\text{and} \qquad y_i = \left(1 + \frac{iDA}{\alpha_{sub}}\right)^{-\alpha_{sub}} \quad , \tag{3}$$

where $P_N(k)$ is the probability of detecting exactly k fault free pages in N pages, M is the required number of fault free pages, y_i is the yield of a cluster of i pages, D is the defect density, α_{sub} is the cluster coefficient, and A is the area of a single page. This simple yield model was chosen, although there are more accurate models available [6]. Nevertheless, the error from uncertainties in circuitry areas and other yield parameters before manufacturing is comparable to the error from using the simpler model. Thus, improving the accuracy of the yield estimation would be difficult. The resulting yield, of a DRAM consisting of n books, Y_{DRAM} can be calculated as the product of $(Y_b)^n$ and Y_{ML}, the total yield of logic circuitry of the memory array, e.g., cache registers and input/output ports.

$$Y_{DRAM} = (Y_b)^n \cdot Y_{ML} \tag{4}$$

No redundancy is applied to these parts of the DRAM. These calculations result in a yield of 99.0% for a 4.0-Mb DRAM when utilizing the described redundancy. Yield without using redundancy can be calculated by applying equation (3) if $i \cdot A$ is replaced by A_{DRAM}, the area of a 4.0-Mb DRAM macrocell. This results in a DRAM yield of 78.7% without redundancy.

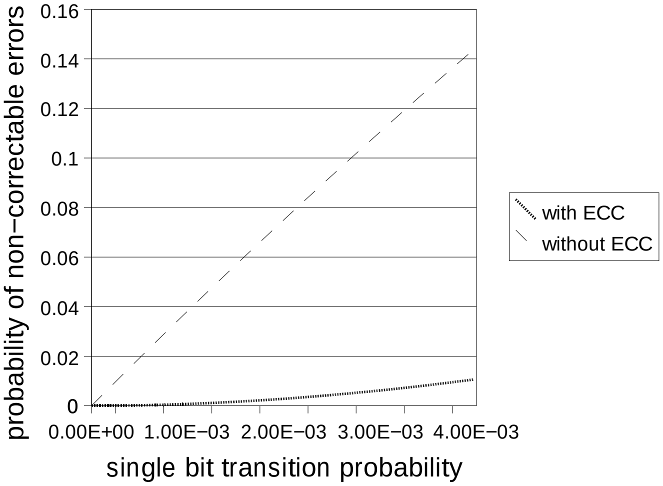

Figure 2. Error correcting capabilities of ECC

4.2. Application of Error Correcting Code

Integrating DRAM in Systems-on-Silicon leads to major advantages concerning area and power consumption in comparison to integrating SRAM. For the system described in this paper, early area estimations showed that 4 Mb DRAM can be integrated with each processing node, compared to 1 Mb when using SRAM. In terms of reliability, intermittent faults have to be considered when utilizing embedded DRAM instead of embedded SRAM.

To address intermittent faults in the DRAM, a (39,32) SEC-DED Hamming encoding/decoding scheme for the memory data has been integrated, which is optimal in terms of logic depth and amount of hardware [7]. In the case of random single-bit failures, the probability P_E of a non-correctable error is given by

$$P_E = 1 - \left[n \cdot p(1-p)^{n-1} + (1-p)^n \right] \tag{5}$$

where p is the transition probability of a single bit and n is the number of bits of a codeword. The assumption of a single-bit failure is true in most cases, as will be discussed later. Figure 2 illustrates the soft-error correction capabilities of the ECC implemented. The probability P_E of a noncorrectable error with ECC is compared to the probability P_{NC} of an error without ECC, which is given by

$$P_{NC} = 1 - (1-p)^m \tag{6}$$

where m is the number of bits of a dataword. Thus, the memory data can be assumed to be error free as long as the single bit transition probability is less than 0.1%.

To implement this coding scheme, the 4-Mb DRAM organized in 4096 pages of 32 words of 32 bits had to be extended. In contrast to designing special DRAM chips, the granularity of the generated DRAM macrocells and the flexibility of the memory controller macrocell were restricted and required an implementation of a total of 5 memory arrays each of 1 Mb organized in 1024 pages of 128 words of 8-bit words and a single 40-bit word memory controller (Figure 3). Thus, a 39-bit code word is distributed over 5 memory arrays and one bit is left unused. This bit is not utilized for implementation of bit line redundancy, since this is covered by the error correcting code. If the bit lines of an internal 64-bit data word are physically interwoven, all multiple-bit failures that do not

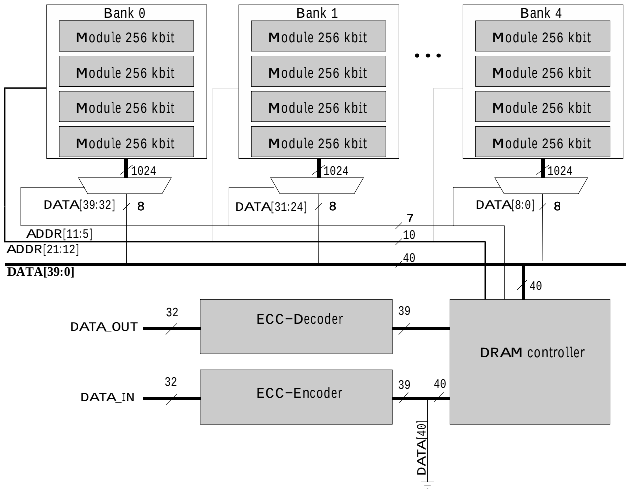

Figure 3. Implementation of error correcting embedded DRAM

exceed 16 fails along a word line (or 8×16 fails in an 8-bit configuration,) can be considered to be single-bit failures of a code word.

The memory overhead is about 37% which is caused by the extra memory array and the five-fold memory caches and ports due to the utilization of five memory banks. Taking into account the total chip area, overhead from the ECC is still 11.6%. Thus, more complex error correcting schemes have not been considered. Area overhead by coding circuitry is negligible,and propagation delay is less than 3 ns for the encoder and less than 6 ns for the decoder. Therefore, DRAM latency is increased by one clock cycle for write and read operations at a clock rate up to 166 Mhz. Due to the area overhead, DRAM yield is decreased to 96.91% if error correction is applied but disabled. By applying the ECC to a memory array manufacturing defects are masked in addition to correcting soft-errors. The effect of the ECC on manufacturing yield is discussed next.

To simplify the calculation, it is assumed that no redundancy is employed. In this case, the yield $Y_{W,ECC}$ of a single word due to ECC is given by equation (7).

$$Y_{W,ECC} = Y_{Cell}^{n} + nY_{Cell}^{n-1}\left(1 - Y_{Cell}^{n}\right) \quad , \tag{7}$$

where n is the number of bits of a codeword and Y_{Cell} is the yield of a single memory cell which can be calculated corresponding to equation (3). Thus, the yield Y_{ECC} of the complete memory array can be given by equation (8).

$$Y_{ECC} = Y_{W,ECC}^{K} \quad , \tag{8}$$

where K is the total number of datawords in the memory.

Arithmetically, this results in Y_{ECC} = 99,99998%.

Even though equation (7) is derived for the case of single-bit errors, the results are also true in the case of multiple-bit errors. Multiple errors that occur along a bit line can be considered to be single-bit errors of several codewords and therefore can be corrected by the ECC. Multiple errors that occur along a word line cannot be corrected by the ECC, but these errors are covered by the word line redundancy of the memory array. Thus, it is essential to test the redundant word lines during the memory test too. Otherwise, the yield of the redundant word lines has to be considered, with the effect being a decrease in the efficiency of reparing multiple-bit errors.

The combination of the ECC and word line redundancy has been proven to have synergistic effects on memory yield in the presence of manufacturing defects [8]. Thus, the resulting yield of the memory array can be assumed to be 100%. The effect of the ECC on area and DRAM yield is summarized in Table 1.

	DRAM area	Overhead	Area of processing node	Overhead	DRAM yield
without ECC	16.88 mm^2	–	42.02 mm^2	–	99.0%
with ECC (disabled)	23.06 mm^2	36.6%	48.78 mm^2	11.60%	96.9%
with ECC (enabled)	23.06 mm^2	36.6%	48.78 mm^2	11.60%	$\sim 100\%$

Table 1. Area and yield of DRAM – impact of ECC

4.3. Yield of a multiprocessor system

The last step of yield calculations is to combine the yield of the embedded DRAM with the yield of the logic circuitry to obtain first the yield of a single processing node and then the yield of the multiprocessor system. No redundancy is provided for the logic circuitry of a processing node since its structure is irregular. Therefore, the yield Y_P of a processing node can be calculated as the product of DRAM yield Y_{DRAM} and logic circuitry yield Y_{logic}.

$$Y_P = Y_{DRAM} \cdot Y_{logic} \tag{9}$$

In the presence of defects, a faulty processing node can be switched off. Thus, the system can still be utilized for video processing tasks, but with lower performance. The yield distribution of such a system is given by

$$P_N(k) = \binom{N}{k} (1 - Y_P)^{N-k} \, Y_P^k \tag{10}$$

where $P_N(k)$ is the probability of getting a system with k working nodes out of N total nodes. This equation considers small area fault clustering, since the area of a processing node is assumed to be larger than the cluster size.

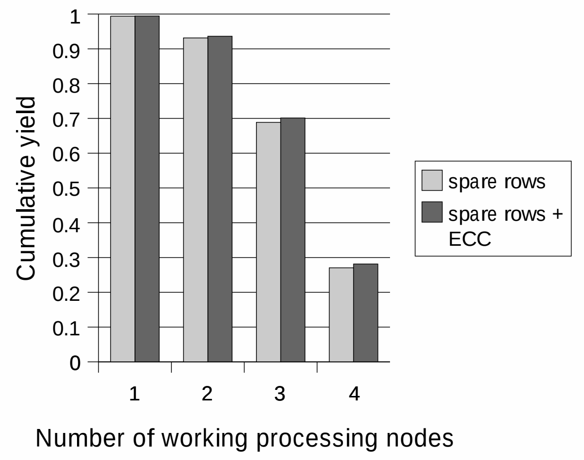

Figure 4. Cumulative yield of multiprocessor system

The yield distribution given by equation (10) is cumulatively plotted in figure 4. The different cases, DRAM without ECC (4 Mb physical), and DRAM with ECC (5 Mb physical), are considered as a parameter of this distribution. The yield distribution is approximately the same, even though the implementation of ECC leads to an area overhead of about 12%. Final synthesis results in a standard cell and SRAM area of 23.00 mm^2 which is comparable to the DRAM area. Nevertheless, the yield of this part of an processing node dominates the yield distribution, because it is not designed to be fault tolerant due to its low degree of regularity. The probability of having a system with four working processor nodes has been predicted to be about 28%.

5. Outlook on implementation of a Large Area Integrated Circuit

Real-time video coding at TV resolution (main profile @ CCIR601 resolution) demands more processing power than can be given by the described multiprocessor system. Performance estimations show that 8 working processing nodes are required to perform these tasks. To maintain the interprocessor bandwidth of the system, the extended multiprocessor system has to be realized on a single piece of silicon. Alternatives, such as MCMs suffer from loss of reliability due to excessive numbers of contacts and due to the Known-Good-Die problem. Extension of the yield estimations of Section 4.3 to the new multiprocessor system leads to an estimated yield of 98% for the 8-processor system if 16 processing nodes are implemented, i.e., 8 extra processing nodes are provided for fault tolerance. The implementation of 12 or 8 processing nodes leads to yields of 79% and 8%, respectively (figure 5).

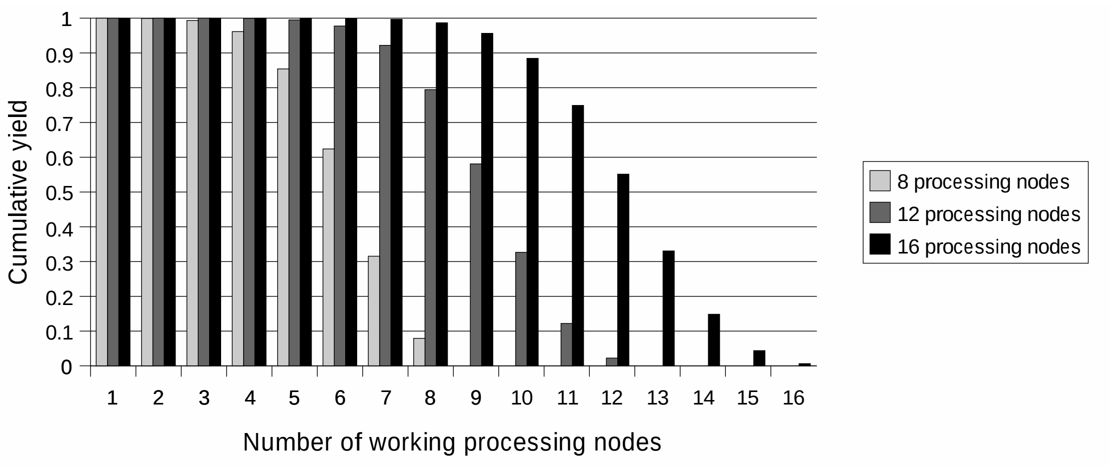

Figure 5. Cumulative yield of Large Area Integrated Circuit

Though the manufacturing process sets an upper limit on the chip area by the size of the reticle, the fabrication of chips of sizes beyond that limit is possible by processing several sets of masks with overlap. Otterstedt et al. presented a method that requires 3 different sets of masks [5]. One set is used for building the processing nodes while the other two provide the interconnect systems as well as input and output pad cells. The main drawback of this method is the cost for the additional sets of mask. Furthermore, handling of different sets of masks is not always possible in the foundry.

To overcome these problems, a methodology has to be developed that allows the manufacturing of Large Area Integrated Circuits by processing a single set of masks with overlap. This methodology has to provide solutions for individualizing the processing nodes, interconnecting the processing nodes over the boundary given by a mask, establishing signal integrity on long interconnects, providing power to the chip and dissipating heat from the chip, connecting the chip with external parts of the system, and housing of these large silicon chips.

This methodology will lead to the realization of compact and reliable high-performance chips (the 100-million-transistor chip [9]) without the necessity to cope with the problems of very deep submicron designs. Candidate systems to take advantage of this methodology feature high-performance and a high degree of parallelism, thus allowing high modularity, which is essential for redundancy on a system level. As this is the case for the multiprocessor system introduced here, the system is a suitable building block of a Large Area Integrated Circuit.

6. Conclusions

In this paper, yield prediction and appropriate yield enhancement measures for a multi processor system with embedded DRAM have been presented. The implementation of such systems requires consideration of both yield loss by manufacturing defects and soft errors in the dynamic memory. It has been shown that the defect tolerance mechanisms of the embedded memory improve the defect tolerance significantly. Thus, the area overhead due to application of an ECC scheme does not decrease the overall system yield. Instead, the ECC enables detection of the dominating soft-errors. Yield predictions show that the realization of Large Area Integrated Circuits is viable with the multiprocessor system as a building block. This results in very compact high-performance systems for high end video coding applications.

Acknowledgments

This work was done in collaboration with the Divisions of Systems Technologies and Semiconductor Technology and Devices at the Laboratory for Information Technology, University of Hannover. The authors would like to thank the other research groups involved in this project. This work is supported by FhG under contract no. T/F41B/T0183/P1307.

References

[1] J. Hilgenstock, K. Herrmann, J. Otterstedt, D. Niggemeyer, and P. Pirsch, “A Video Signal Processor for MIMD Multiprocessing,” Design Automation Conference (DAC), 1998, pp. 50-55.

[2] J. Hilgenstock, K. Herrmann, and P. Pirsch, “Memory Organization of a Single-Chip Video Signal Processing System with Embedded DRAM,” Great Lakes Symposium on VLSI, 1999, pp. 42-45.

[3] *Coding of Moving Pictures and Audio, Overview of MPEG-4 Profiles and Levels*, ISO/IEC JTC1/SC29/WG11, MPEG98/N2325, July 1998.

[4] B. Könemann, J. Mucha, and G. Zwiehoff, “Built-In Test for Complex Digital Integrated Circuits,” IEEE J. Solid-State Circuits, Vol. 15, No. 3, 1980, pp. 315-318.

[5] J. Otterstedt, K. Gaedke, K. Herrmann, M. Kuboschek, H. U. Schröder, and A. Werner, “A 16.6 cm^2 Monolithic Multi Processor System Integrating 9 Video Signal-Processing Elements,” International Solid State Circuit Conference 1996, Digest of Technical Papers, pp. 306-307, 464, Slide Supplement to Digest of Technical Papers, pp. 242-243, 450.

[6] I. Koren, Z. Koren, and C. H. Stapper, “A Unified Negative-Binomial Distribution for Yield Analysis of Defect-Tolerant Circuits,” IEEE Trans. Computers, Vol. 42, No. 6, 1993, pp. 724-734.

[7] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice Hall, 1983.

[8] C. H. Stapper and H.-S. Lee, “Synergistic Fault-Tolerance for Memory Chips,” IEEE Trans. Computers, Vol. 41, No. 9, 1992, pp. 1078-1087.

[9] L. Geppert, Y. Taur, B. Chappel, N. Harned, L. R. Harriot, D. Herrel, and Y. Zorian, “The 100-million-transistor IC,” IEEE Spectrum, July 1999, pp. 22-60