Hierarchical Fault Tolerance Memory Architecture With 3-Dimension Interconnect

Da Wang1,2, Yuanjiang Xie1,2, Yu Hu1, Huawei Li1, Xiaowei Li1
1 Key Laboratory of Computer System and Architecture, Institute of Computing Technology, Chinese Academy of Sciences, P.O. Box 2704, Beijing, China, 100080
2 Graduate University of Chinese Academy of Science, Beijing China, 100049
E-mail: {wangda, xieyuanjiang, huyu, lihuawei, lxw}@ict.ac.cn

Abstract - This paper proposed hierarchical fault tolerance techniques for ultrahigh-density memories based on 3-dimension interconnect technology. It describes how to implement hierarchical architecture with different granularity redundancies and how to combine error correction code (ECC), built-in self-test (BIST), built-in repair-analysis (BIRA), and built-in self-repair (BISR) capabilities. Simulation is employed to estimate the memory behavior of various configurations, and experimental results indicate that the proposed method has substantial reliability improvements over conventional techniques. For a memory with 1% bit-level failure rate and 50% device-level defect density, the proposed method can gain 100% reliability by using less than 30% extra overhead. It proves the availability of the proposed architecture.

I. INTRODUCTION

The unremitting requirement of smaller and faster electronics brings tremendous challenges in CMOS dominating semiconductor industry. Moore’s law states that not only the speed and performance of integrated circuits (ICs) roughly doubles every 18 months, but also the density of ICs doubles every two years. As the technology is quickly approaching a brick wall, 3-dimension interconnect (3D for short) technology is likely to be one of the best potential solutions to extend Moore's Law [1-5]. Though a real IC design proves the advantages of the 3D technique [4], there are still challenges on new approaches for designing, fabricating, and testing. Since electronic device geometries keep on shrinking [5], the inherent inexactness of fabrication technique for 3D structure will decrease IC designs reliability sharply even in regular structures such as memory arrays. Although memory bit-level error rates tend to decrease with each new generation of technology, huge increases in bit density result in device-level defect densities reach to the level of 5%–10% [6-9], which requires more powerful fault tolerance and fault avoidance measures.

To the best of our knowledge, this is the first work on fault tolerance methods for 3D memories. Our work has two major contributions: firstly, it proposes hierarchical fault tolerance architecture for 3D memories. Secondly, it analyzes the reliability of the proposed architecture in the combination of self-repairing and ECC schemes. The rest of this paper is organized as follows. Section II introduces the main background of this paper, including a brief overview of 3D technology our work based and conventional fault tolerance methods. Section III illustrates the proposed memory architecture and implementation strategies in detail. Section IV analyzes the capability of this method to improve 3D memory reliability and gives experiment results. Section V concludes the paper.

II. PRELIMINARY

A. 3-Dimention Technology

3D techniques have been studied for more than ten years to address various issues [1-4]. The basic concept of 3D is to integrate planar circuit layers into 3-Dimention with short vertical interconnects as shown in Figure 1. Further more, 3D technology is also suitable for manufacturing to integrate heterogeneous circuits together. For example, it enables integration of multi-function circuits in a same chip, such as analog logic, digital logic, memories, and communication applications, as shown in Figure 1. With the simple illustration given in Figure 2, it shows that the wire length of 3D ICs is much shorter than that of 2Ds’. Researches in [2-3] analyzed the performance advantages of 3D comparing to that of traditional 2Ds’ by applying Rent’s rules. The researches reported that 3D structure could enhance the IC performance efficiently.

Figure 1. Conceptual view of mapping planar to heterogeneous 3D IC

Figure 2. Comparison of longest path in 2D and 3D designs [2]

With Via-first/Face-down/Cu-Cu bond technique, a 3D die stacking implementation is applied to an Intel deeply pipelined high performance microprocessor [4], iA32a, and the processor floorplan in planar and 3D states are shown in Figure 3. It also reports that the 3D implementation can potentially improve 15% performance at the cost of 15% more power dissipation. Though thermal dissipation and die-to-die via density are disadvantages of 3D structures [2-4], the benefits brought by this technique are obvious and ITRS has identified that 3D interconnect is one way to achieve
more improvement in IC performance [5]. It is necessary to study fault tolerance methods for 3D designs required to be high reliable.

![Figure 3. Floorplan of iA32 microprocessor [4](a) Planar floorplan; (b) 3D floorplan](image)

B. Fault Categories and Fault Tolerance Methods

In general, faults detected by semiconductor devices are fallen into three main categories [6]: 1) Permanent faults, which lie on irreversible physical defects; 2) Intermittent faults, which occur repeatedly at the same location because of unstable or marginal hardware, and tend to occur in bursts when the fault is activated by environmental changes, such as temperature or voltage fluctuation; 3) Transient faults, which are runtime errors introduced by neutron and alpha particles radiation, electromagnetic interference, electrostatic discharge, or power supply and interconnect noises.

Fault tolerance is widely used to increase the reliability of VLSI designs. It is implemented at circuit or system level, and relies on methods such as error detection, error recovery, redundant elements remapping, timing redundancy, and ECCs [7-11]. BIST, BIRA and BISR circuitries are usually used for detecting faults and remapping uncorrectable memory cells to fault-free ones. Nowadays, though intermittent faults occur more frequently than permanent faults, fortunately, similar to permanent faults, these faults can also be easily removed by remapping process. ECC methods are used to tolerant transient faults which are runtime errors unable to be removed by redundancies.

III. PROPOSED HIERARCHICAL FAULT TOLERANCE MEMORY ARCHITECTURE

A. Architecture Overview

The considered memory architecture is based on a 3D die stacking structure with hierarchical fault tolerant elements. As shown in Figure 4, level-0 memory is the basic memory component and is comprised of memory arrays, decoding logic, private redundant rows and columns, as well as level-0 fault replace and correct unit (FRCU-0). It is the only memory using both remapping and ECC methods. The memory array is the main storage body of level-0 memory. Decoding logic is used for address decoding and restructuring. Private spare rows/columns are only used to replace faulty cells of its host level-0 memory array. The FRCU-0 is a processing unit which can dynamically provide fault tolerance functions for its level-0 memory.

Level-1 memory logically consists of level-0 memories, and is addressed by the level-1 FRCU, FRCU-1. FRCU-1 controls the public redundant rows and columns, as shown in Figure 5. These spares are used to replace remained uncorrected circuits in level-0 memories after the FRCU-0 processing step.

![Figure 4. Floorplan of Level-0 Memory](image)

Die-0 is a memory die consists of many level-1 memories. Since 3D structure is easy to put different dice together, to trade-off memory performance and reliability, the spare elements and FRCU logic for the overall die-0 level is set in another die. For example, public redundant elements in die-1 which is close to die-0 are spares and FRCU logic for die-0, as depicted in Figure 5. A complete hierarchical design for the proposed architecture is given in Figure 5. Analysis in this paper is based on this frame.

C. Fault Tolerance Techniques Used in the Proposed Architecture

Though memory bit-level errors tend to fall, its defect density in device-level is kept on increasing. Efficacious fault tolerance method is needed in such situation. With the trade off between area and reliability, the fault tolerance method used in this paper is based on remapping and ECC functions.
Level-0 memory is the only memory using both remapping and ECC methods. As discussed in [12], the relationship between data bits $k$ and check bits $r$ in a word which using single-error-correcting and double-error-detecting (SEC-DED) coding is given in:

$$k \leq 2^{r-1} - r.$$  \hspace{1cm} (1)

For example, 7 extra check bits are needed for a 32 data bits word. With considering area cost, SEC-DED is chosen as ECC method in this paper. Remapping is done with redundant rows and/or columns. The latency of an optimal remapping structure is supposed bearable by small size memory. These two fault tolerance functions are controlled by local FRCU-0 logic. Firstly, FRCU-0 will run BIST to detect errors, and then analysis the result, then it will execute SEC to correct all single bit errors in the words. After SEC, the strategy will rerun BIST to detect uncorrected faults. If there are still faults, the strategy will remap memory structure with the most-repair algorithm. In the case that the memory remained faulty after using all level-0 spares, FRCU-0 will report the remained errors to higher level FRCU logic, FRCU-1, demanding for more spares.

2) Fault tolerance in level-1 memory:

With the memory floorplan unknown, only redundancy remapping method is used in this level. FRCU-1 is the fault tolerance processing unit in level-1 memory. It runs BIST to detect errors, gathers information from FRCU-0s, and gives optimal resource assignment to implement remapping for remained faulty rows and/or columns of local level-1 memory. After the processing of FRCU-1, if there is still uncorrectable level-0 memories in die-0, FRCU-1 will report the faulty level-0 memories information to the next higher level FRCU processing unit which is located in die-1.

3) Fault tolerance in public redundant part:

The components of this level are totally different from the others. Level-0 memories are used as public spare elements to swap faulty elements in die-0. They are under control of FRCU-2, which processes all failure messages coming from FRCU-1s and finding a solution to gain reliability as much as possible for the memory of die-0. The proposed hierarchical fault tolerance memory architecture can be extended beyond more levels of redundancy and not tied to specific technology. The architecture of this paper is proved available in next section.

IV. EXPERIMENT ENVIRONMENT AND RELIABILITY ANALYSIS

A. Experiment Setup

A simulator is developed to model the performance of large random access memories by randomly injecting permanent, intermittent, and transient faults with their location and duration throughout the usable storage bits and ECC bits of the memory. The FRCU logic processes errors detected by BIST with following requests:

1) If only one-bit error occurs in data bits and ECC bits of the word, the word is reported being corrected by ECC method without asking for spares.

2) If more than one-bit error detected in a word during the first time running of BIST, it will be reported as a failure word needing further test.

3) In the diagnosis of FRCU, if the failure word reported in the 2) step diagnosed as a hard error, it will be swapped by a spare element directly, if it reported as a soft one, FRCU will ignore it.

The simulator simulates the runtime responses of FRCU logics, and calculates the repair rate. Since it is better to set level-0 and level-1 memory as small as possible to reduce impact on memory performance, we injected 1% bit-level failure rate and 50% device-level defect density for various sizes of memories, and then, obtained the reliability experimental results under different spare ratios. For simplicity, we assume all redundant elements are fault free in the experiments, although the redundant elements in real chips may be faulty. In our future work, we will take the faulty redundant elements into consideration.

B. Reliability Analysis

The set of curves shown in the following figures are based on a multilevel memory which has 32MB usable
storage cells in the die-0. The 32MB cells are abstracted to 256 level-1 memory arrays. In our experiments, we set different sizes for level-0 memories. The curves in Figure 6 depict the relationship between extra cost ratio and the bit-level errors ratio when achieving 100% reliability under 50% device-level defect density. The size of level-0 memories are 1024×32, 2048×32, 1024×64, 512×128, and 256×256. For each level-0 memory, only one spare column is set, but spare rows are setting changeable. For level-1 memories, the number of public redundant column is also one and the number of redundant rows is a variant. At the same time, there is only one level-0 memory in die-1 as spare element. According to formula (1) the SEC-DED coding cost for a 32 data bits word is 7 bits, for a 64 data bits word is 8 bits, for a 128 data bits word is 9 bits, and for a 256 data bits word is 10 bits.

Figure 6 (a) shows the area overhead on ECC bits and redundant memory elements. We can see the 64 data bits word memory gains the most benefit in the case of high defect rate. Because its rows and columns proportion is close to the overhead proportion on ECC bits and spares elements. Figure 6 (b) compares two memories that have different usable storage size but same data bits word of level-0 memory. From the experimental results, it is clear that if the rows and columns proportion is large, smaller level-0 memory can gain high reliability with less extra cost. For fixing up only one spare column for level-0 memory, in Figure 6(c), it shows that 256×256 memory needs more extra cost to gain 100% reliability though its ECC coding cost is the least. So how to effectively set memory structure and spares needs further study. The experiment results verified that the proposed method is useful to enhance 3D memory reliability with tolerable cost.

V. CONCLUSION

The goal of this paper is to study methods for improving the reliability of 3D interconnect memories through a novel hierarchical redundant architecture, which is implemented with the combination of ECC, BIST, and BISR capabilities. The proposed method is believed not only executing at runtime, but also detecting and correcting the faults before they result in uncorrectable errors. Though the experiment is not conducted on a real fabricated memory, from our experiment results, we believe that the proposed method is viable to enhance the memory reliability for 3D memories even with high defect density. A lot of work is still needed to consummate the memory architecture study for obtaining high reliability in future. It is our hope that this paper could generate renewed interest in the research community towards this field.

REFERENCES