Microelectronics Reliability 61 (2016) 43–47

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/mr

A cross layer approach for efficient thermal management in 3D stacked SoCs

Matthias Jung *, Christian Weis, Norbert Wehn

University of Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 15 September 2015
Accepted 14 December 2015
Available online 28 December 2015

Keywords:
DRAM
Retention
Refresh
3D-integration
Reliability
Approximate computing
Error resilience

ABSTRACT

3D stacking of silicon dies via Through Silicon Vias (TSVs) is an emerging technology to increase performance, energy efficiency and integration density of today's and future System-on-Chips (SoCs). Especially the stacking of Wide I/O DRAMs on top of a logic die is a very promising approach to tackle the memory wall and energy efficiency challenge. The potential of this type of stacking is currently under investigation by many research groups and companies in particular for mobile devices. There, for instance, the baseband processing and the application processor can be implemented on the same single logic die. On top of this die one or several Wide I/O DRAMs are stacked. An example of such a SoC is the WIOMING chip [15]. However, new challenges emerge, especially thermal management, which is already a very demanding challenge in current 2D SoCs. With 3D SoCs this problem exacerbates due to reliability issues such as the temperature sensitivity of DRAMs, i.e., the retention time of a DRAM cell largely decreases with increasing temperature.

In this paper, we show a holistic cross layer reliability approach for efficient reliability management starting from measuring and modeling of DRAM retention errors, which finally leads to optimizations for specific applications. These optimizations exploit the data lifetime and the inherent error resilience of the application, which is for instance given in the probabilistic behavior of wireless communications.

1. Introduction

The increased power density and thermal dissipation of today's processing and memory systems restrict application performance on handheld devices like smartphones as well as on high-end servers. Advanced fabrication processes that use 3D packaging based on TSV technology enable even tighter integration of systems in a small form factor. These 3D systems start to break down the memory and bandwidth walls. However, this largely increases the power density and reduces the heat dissipation properties of the aggressively thinned dies to enable reliable TSV production. In fact, a 3D stacked SoC aggravates the thermal crisis, which can provoke errors in circuits. This is especially important for *Dynamic Random Access Memories* (DRAMs) as they are highly sensitive to temperature changes, and have to be refreshed regularly due to their charge-based bit storage property (capacitor). Due to the much increased leakage at the cells the refresh frequency needs to be adjusted accordingly to avoid retention errors [10,13] as shown in Fig. 1.

The retention time of a DRAM cell is defined as the amount of time that a DRAM cell can safely retain data without being refreshed [5]. This DRAM refresh operation, as already mentioned, must be issued periodically, and causes both performance degradation and increased energy consumption, both of which are expected to worsen as the DRAM density increases (almost 50% of future DRAMs total energy, as shown in [6]). Not all DRAM cells behave in the same leaky way, due to process variations. Many prior studies assume that it is possible to keep track of relatively few weak DRAM cells (low retention time) [7,8] and therefore to reduce the impact of refreshing by avoiding the usage of these cells. However, the effects of data pattern dependence and variable retention time (VRT), which we also observed during our measurements, inhibit the application of DRAM retention time profiling mechanisms. Nonetheless, Liu et al. proposed in [9] an unreliable and a reliable region in the DRAM and refreshed them at different rates.

To tackle the aforementioned challenges, advanced error and thermal modeling [1] together with high-level simulators, such as gem5 [2] and DRAMSys [3,4], are required to investigate system behavior with advanced MPSoCs based on DDRX or WIDE I/O DRAMs. In contrast to [9] we exploit in our work the feasibility to omit refreshing the DRAM [10], for dedicated applications.

The major contributions of this work are:

1. We measure the retention times and provoked bit errors of WIDE I/O DRAM dies on top of a SoC logic die (see Section 3).
2. We propose a calibrated DRAM retention error model based on the measurements, which can be integrated into simulators like gem5 [2], DRAMSim2 [11] and DRAMSys [3,4] (see Section 4).
3. With this model we show that dedicated applications can tolerate DRAM retention errors: Either the lifetime of the data is shorter than the currently required DRAM refresh period, or the application

* Corresponding author at: University of Kaiserslautern, Paul-Ehrlich-Strasse 12, 67663 Kaiserslautern, Germany.

E-mail address: jungma@eit.uni-kl.de (M. Jung).

http://dx.doi.org/10.1016/j.microrel.2015.12.025

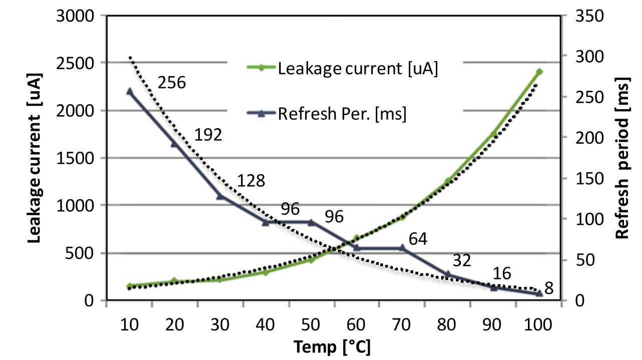

Fig. 1. Leakage current and required refresh periods at different temperatures [10,13].

can tolerate bit errors to some degree in a given time window (see Sections 2 and 6).

2. Applications with inherent error resilience

Applications differ in their characteristics of data lifetime, storage and size. Due to iterative processing and inherent error resilience, we selected two representative examples for our investigations. The first example is a graph processing application in a Big Data environment and the second example is a channel decoding application which is part of any wireless baseband processing.

2.1. Graph processing

Complex graphs are at the heart of today's big data challenges like recommendation systems, customer behavior modeling, or incident detection systems. One recurring task in these fields is the extraction of network motifs, recurring and statistically significant subgraphs.

In [14] an efficient method is presented for calculating similarities in large graphs using the metric of co-occurrence: Fig. 2 explains the basic ideas: The similarity between *you* and *Liam* is based on the number of common friends, the so-called co-occurrence: $coocc(you; Liam) = 3$. The question arises whether the number three is significant. Assume *you* and *Liam* have thousands of friends, versus you have only three. For this random graphs are created based on the same degree sequence and the premise that nodes have no similarities. To get the random graphs a sufficient number of pairs of edges are swapped, drawn uniformly at random, if and only if no multiple edges would arise due to the swap. This generates independent graphs with the same degree sequence, see Fig. 2. The expected co-occurrence can be calculated by generating many of these graphs. That information can judge how significant the similarity in the original graph is.

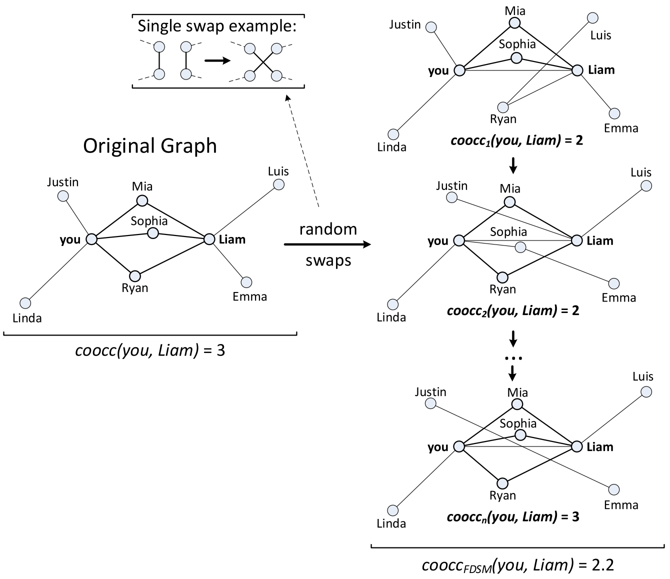

Fig. 2. Co-occurrence and swapping in a graph [14,10].

However, for large graphs this is in general a very time and memory consuming task on standard computing clusters. Netflix, a commercial video streaming service, has released 100,480,507 user ratings for all of their 17,700 movies from 480,189 users [21]. Assuming one co-occurrence operation per clock cycle at 1 GHz would require 115 days to evaluate this data set. An optimized accelerator ASIC has been presented in [14], where the graph is stored in the DRAM as sparse adjacency matrix. Due to the stochastic behavior, it is very likely that a certain amount of bit errors in the large graph matrix will not influence the quality of the result. We will confirm this assumption by simulations in Section 6.

2.2. Channel decoding

In digital communication systems, channel coding plays an integral role in all important modern communication standards. One of the currently most sophisticated concepts of channel coding is *Low-Density Parity Check* (LDPC) coding, which is part of recent wireless communication standards like WiGig and WiFi. The strength of LDPC coding is its iterative decoding algorithm that allows the utilization of information on the reliability of each bit, so-called soft-information, which largely improves the communications performance. Investigations have shown that channel decoding algorithms have an inherent error resilience. Gimmler et al. present in [22] that errors in the memories can be tolerated up to a certain degree. Thus, if the data are stored in the DRAM it is very likely that the influence of DRAM retention errors on the communications performance is negligible, as we will show in Section 6.

3. Measurements with WIOMING chip

The WIOMING 3D-IC [15] is an SoC with a stacked Wide I/O DRAM, similar to [16]. Using this 3D-IC we conduct a set of experiments [1] to measure the retention time and bit error behavior of WIDE I/O DRAMs. For this paper, we extended the measurements with respect to a higher dynamic range of retention times. We executed the tests with the WIOMING chip multiple times to manifest the results. Fig. 3a shows the measurement results. We clearly see in the plot the data pattern dependency (DPD), similar to the study in [12] for commodity DRAMs. The reason for this phenomenon is bitline–bitline, bitline–cell and bitline–wordline coupling.

Moreover, we observe that bits are only flipping from “1” to “0”. This fact is because of this WIDE I/O DRAM uses true-cells, which means that the data value is stored as it is. Other DRAM vendors often use an anti-cell implementation where for half of the DRAM cells the inverse value is stored [12].

In contrast to [17] we observe that the majority of cells in this WIDE I/O DRAM device can hold data much longer than 10,000 s for a 0xFF data pattern, shown in Fig. 3b. After 10,000 s only 6% of the cells are flipped for 40 °C while 28% are flipped for 105 °C.

Additionally, we see in Fig. 4 the effect of *Variable Retention Times* (VRTs) [12] and the uniform distribution for weak DRAM cells. Not all cells are permanently failing beginning from a specific temperature. For instance, the red triangle (90 °C) fail shown in the circle is a typical VRT fail, as it disappears at higher temperature. More detailed analysis and results as well as further measurements with DDR3 devices can be found in [1].