

Reliability and Thermal Challenges in 3D Integrated Embedded Systems

Christian Weis, Matthias Jung and Norbert Wehn
University of Kaiserslautern, Germany

Abstract—Tightly coupled embedded system processors and accelerator IPs with memories, such as DRAM or NVM, using 3D integration technology based on TSVs (Through Silicon Vias) offer high bandwidth memory access at low energy consumption. However, those memories, especially DRAMs, are very sensitive to temperature changes since they use capacitors as volatile bit storage elements. The 3D stacking of these layers increases the challenges, such as high power densities and thermal dissipation. It also has a much stronger impact on the retention time of 3D stacked WIDE I/O DRAMs that are mostly placed on top of the heterogeneous embedded processing system (MPSoC). Consequently, it is very important to study the temperature and retention behavior of WIDE I/O DRAMs. Furthermore, there is the need to explore on high-level with advanced holistic modeling how reducing or omitting the refresh in the DRAM influences executed applications. Additionally, it is beneficial to quantify these effects and impact on standard FPGA platforms using commodity DDR3 SO-DIMMs. The outcomes of these investigations permit the transfer of properties and parameters to more advanced 3D stacked systems.

In this work, we demonstrate the impact of cross-layer DRAM refresh policies on three selected applications ranging from image processing on FPGA, baseband wireless processing with a LDPC-decoder, to a big data application using co-occurrence processing for large graphs. Our results highlight the inherent application resilience wrt. DRAM retention time errors (bit flips) for these very diverging applications.

I. INTRODUCTION

The increased power density and thermal dissipation of today's processing and memory systems restrict application performance on 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 increases largely 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 DRAMs as they are highly sensitive to temperature changes, and have to be refreshed regularly due to their charge-based bit storage property (capacitor). Thus, advanced error and thermal modeling together with high-level simulators, such as gem5 [1] and DRAMSys [2], are required to investigate current or future processing and main memory systems (based on DDR3 or WIDE I/O DRAMs).

Furthermore, 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 [3]. This DRAM refresh operation must be issued periodically, and causes both performance degradation and increased energy consumption (almost 50% of future DRAMs total energy), both of which are expected to worsen as the DRAM density increases [4]. 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) [5], [6] and therefore reducing 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 measurements, inhibit the application of DRAM retention time profiling mechanisms.

Nonetheless, Liu et al. proposed in [7] an unreliable and a reliable region in the DRAM and refreshed them at different rates. In contrast to [7] we exploit in our work the feasibility of using a *no refresh policy* on three dedicated applications. These three applications are as follows:

- First, we demonstrate the influence of disabling the refresh completely for an image processing task on a XILINX FPGA rotating the image in the DRAM (application specific memory controller).
- Second, we present the input data resilience of a LDPC decoder IP for wireless baseband applications.
- Third and finally, we show for a big data application, namely the co-occurrence calculation based on large data graphs (recommendations for social networks, netflix and more), the impact of the *no refresh policy* on the quality of the prediction.

The results of this work clearly highlight the feasibility of using a no refresh policy on selected applications that can tolerate a certain level of bit error probability (bit flips based on retention time errors in the DRAM). Exploiting this possibility is a viable path for achieving less power consumption and higher available memory bandwidth in a given embedded computing system. The presented investigations are work-in-progress and the results are currently not published.

REFERENCES

- [1] Nathan Binkert, et al. *The gem5 simulator*. SIGARCH Comput. Archit. News, 39(2):1–7, August 2011.
- [2] Matthias Jung, et al. *TLM modelling of 3D stacked wide I/O DRAM subsystems: a virtual platform for memory controller design space exploration*. In Proceedings of the 2013 Workshop on Rapid Simulation and Performance Evaluation: Methods and Tools, RAPIDO '13, pages 5:1–5:6, New York, NY, USA, 2013. ACM.
- [3] T. Hamamoto, et al. *On the retention time distribution of dynamic random access memory (DRAM)*. Electron Devices, IEEE Transactions on, 45(6):1300–1309, Jun 1998.
- [4] Jamie Liu, et al. *RAIDR: Retention-Aware Intelligent DRAM Refresh*. In Proceedings of the 39th Annual International Symposium on Computer Architecture, ISCA '12, pages 1–12, Washington, DC, USA, 2012. IEEE Computer Society.
- [5] J.-H. Ahn, et al. *Adaptive Self Refresh Scheme for Battery Operated High-Density Mobile DRAM Applications*. In ASSCC. IEEE Asian, Nov 2006.
- [6] R.K. Venkatesan, et al. *Retention-aware placement in DRAM (RAPID): software methods for quasi-non-volatile DRAM*. In Proc. of HPCA, 2006.
- [7] Song Liu, et al. *Flicker: Saving DRAM Refresh-power Through Critical Data Partitioning*. SIGPLAN Not., 46(3):213–224, March 2011.