How an Undervolted High-Bandwidth Memory Looks Like

Seyed Saber NabaviLarimi *†,1

* *Barcelona Supercomputing Center (BSC), Barcelona, Spain*
† *Universitat Politecnica de Catalunya (UPC), Barcelona, Spain*

ABSTRACT

High Bandwidth Memory (HBM) has been introduced as a solution to DRAMs' bandwidth and power inefficiency and has enjoyed wide industry adoption in recent years. Since this memory is integrated into a single package along with computation chips, it will take a bite out of the package's power budget. To remedy this, we can use undervolting (also called voltage underscaling). The downside of undervolting is that it can cause a malfunction in some parts of the device. In this work, we will present a high-level view of how the fidelity of HBM changes as we decrease its supply voltage.

KEYWORDS: HBM; Undervolting; Fault-Map

1 Introduction

With the rapid growth in the computational capacity of modern systems, Dynamic Random Access Memory (DRAM) has become a bottleneck in terms of performance and power efficiency, especially for data-intensive applications. Since these problems are rooted in the fact that data transfer to/from DRAM is inefficient, improving the memory architecture in a way that improves data transfer can be the solution. Some of these new architectures include Reduced Latency DRAM (RLDRAM), Graphics DDR (GDDR), and Low-Power DDR (LPDDR). Along the same line, High-Bandwidth Memory (HBM) [Kim14] is invented to bridge the *bandwidth* gap between computing devices and their main memory.

Manufacturing HBM devices requires stacking multiple DRAM layers and placing that stack next to the computing elements in the same package. This integration will improve bandwidth, power consumption and reduces the form factor. The downside is that the computing elements now have to share their power budget with the main memory. To overcome this obstacle, we implement Undervolting, also called voltage underscaling. This technique reduces supply voltage but leaves the frequency intact, thereby saving power without negatively affecting performance. The limitation to undervolting is that if we push the voltage too low, some parts of the system might stop working. In this poster session, we will show

[1]E-mail: saber.nabavi@{bsc.es, upc.edu, gmail.com}

how much power we can save with undervolting and how HBM behaves if we push its supply voltage too low.

References

[Kim14] Joonyoung Kim; Younsu Kim. HBM: Memory Solution for Bandwidth-Hungry Processors. In *HCS*, 2014.