H3DM: A High-bandwidth High-capacity Hybrid 3D Memory Design for GPUs

NEGAR AKBARZADEH, Sharif University of Technology, Iran
SINA DARABI, Institute for Research in Fundamental Sciences (IPM), Iran, Switzerland
ATIYEH GHEIBI-FETRAT, Sharif University of Technology, Iran
AMIR MIRZAEI, Sharif University of Technology, Iran
MOHAMMAD SADROSADATI, Institute for Research in Fundamental Sciences (IPM), Iran
HAMID SARBAZI-AZAD, Sharif University of Technology & Institute for Research in Fundamental Sciences (IPM), Iran

Graphics Processing Units (GPUs) are widely used for modern applications with huge data sizes. However, the performance benefit of GPUs is limited by their memory capacity and bandwidth. Although GPU vendors improve memory capacity and bandwidth using 3D memory technology (HBM), many important workloads with terabytes of data still cannot fit in the provided capacity and are bound by the provided bandwidth. With a limited GPU memory capacity, programmers should handle the data movement between GPU and host memories by themselves, causing a significant programming burden. To improve programming ease, GPUs use a unified address space with the host that allows over-subscribing GPU memory, but this approach is not effective in terms of performance once GPUs encounter memory page faults.

Many recent works have tried to remedy capacity and bandwidth bottlenecks using dense non-volatile memories (NVMs) and true-3D stacking. However, these works mainly focus on one bottleneck or do not provide a scalable solution that fits future requirements. In this paper, we investigate true-3D stacking of dense, low-power, and refresh-free non-volatile phase change memory (PCM) on top of state-of-the-art GPU configurations to provide higher capacity and bandwidth within the available area and power budget. The higher density and lower power consumption of PCM provide higher capacity through integrating more cells in each 3D layer and enabling stacking more layers. However, we observe that stacking more than six layers of pure-PCM memory violates the thermal constraint and severely harms the performance and power efficiency due to its higher write latency and energy. Further, it degrades the lifetime of GPU to less than one year.

Utilizing a hybrid architecture that leverages the benefits of both DRAM and PCM memories has been widely studied by prior proposals; however, true-3D integration of such a hybrid memory architecture especially on top of state-of-the-art powerful GPU architecture has not been investigated yet. We experimentally demonstrate that by covering 80% of write requests in DRAM and eliminating refresh overhead, true-3D stacking of eight 32GB layers of PCM along with two 8GB layers of DRAM is possible resulting in a total of 272GB memory capacity. Based on the explored design requirements, We propose a 3D high-bandwidth high-capacity hybrid memory (*H3DM*) system utilizing a hybrid-3D (H3D)-aware remapping scheme to reduce expensive PCM

Authors' addresses: Negar Akbarzadeh, n.akbarzadeh93@gmail.com, Sharif University of Technology, Tehran, Tehran, Iran; Sina Darabi, sinad1367@gmail.com, Institute for Research in Fundamental Sciences (IPM), Iran, Lugano, Tcino, Switzerland; Atiyeh Gheibi-Fetrat, atiye.gheibi@gmail.com, Sharif University of Technology, Tehran, Tehran, Iran; Amir Mirzaei, amir.mirzaei1379@gmail.com, Sharif University of Technology, Tehran, Tehran, Iran; Mohammad Sadrosadati, msadr89@gmail.com, Institute for Research in Fundamental Sciences (IPM), Tehran, Tehran, Iran; Hamid Sarbazi-Azad, azad@sharif.edu,azad@ipm.ir, Sharif University of Technology & Institute for Research in Fundamental Sciences (IPM), Tehran, Tehran, Iran.

2476-1249/2024/3-ART12 $15.00
https://doi.org/10.1145/3639038