Performance evaluation of High Bandwidth Memory for HPC Workloads

Amit Kumar Kabat
Electrical Engineering Department
Indian Institute of Technology Madras
Chennai, India
ee20m071@smail.iitm.ac.in

Shubhang Pandey
Electrical Engineering Department
Indian Institute of Technology Madras
Chennai, India
ee19s057@smail.iitm.ac.in

Venkatesh Tiruchirai Gopalakrishnan
Electrical Engineering Department
Indian Institute of Technology Madras
Chennai, India
tgvenky@ee.iitm.ac.in

2022 IEEE 35th International System-on-Chip Conference (SOCC) | 978-1-6654-5985-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/SOCC56010.2022.9908071

***Abstract*—Recent advances in 3D integrated fabrication have allowed the development of 3D stacked memory. The technology presents itself as a viable solution to the memory wall problem. The 3D memory has stacked DRAM layers over the logic die connected through Through Silicon Vias (TSVs). In this paper, we study the performance of latency and energy of one such 3D stacked memory, namely the High Bandwidth Memory (HBM). We quantify the performance improvement of HBM2 against its predecessor technology, GDDR5, and competing mature technology, GDDR6. We have integrated the DRAMSim3 simulator with the Structural Simulation Toolkit simulator and have studied the High-Performance Computing Workloads from Rodinia Benchmarks Suite (RBS). In our paper, we carry out the evaluation of HBM2 and GDDR6 technology by varying the following metrics of the synthetic benchmarks- the number of instructions, and the read to write ratio. We observe the total execution time, memory cycles, and energy consumed for the RBS benchmarks. Our evaluation for synthetic benchmarks reveals that the read write latency performance of HBM2 is the orders of magnitude better than the GDDR6 technology. For real-world benchmarks, the execution time for HBM2 provides $\sim 30\%$ improvement compared to GDDR6 and energy savings of $\sim 23\%$ from GDDR6 to HBM2.**

***Index Terms*—High Bandwidth Memory, GDDR5, GDDR6, HPC Workloads**

I. Introduction

In a conventional computer, data and instructions are kept in the main memory, and the CPU uses an external data bus to access them to perform computation. Over the years, the processor speed has increased exponentially, while the improvement in memory access time has not been that rapid. This issue is known as the memory wall problem [1]. Architectural solutions developed in the past decade such as pipelining, on-chip cache, multicore systems, and many more have constantly improved the performance of the processor thereby overshadowing the memory wall problem [2]. Nevertheless, it has always been quite evident that the system will hit a performance-limiting wall due to memory badwidth (BW) requirement and access speed inspite of all these solutions. The current real-world applications from deep learning, image processing, graph processing, and data analysis are highly data-intensive and will only become more and more intensive over the years. So new approaches to the computing system are needed to tackle it.

There is no single obvious solution to the memory wall problem. Overall, there are two issues to be addressed- first

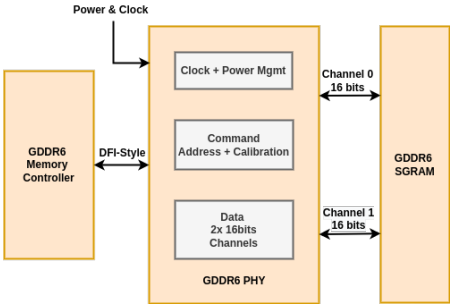


Fig. 1. GDDR6 memory with interface

is the time taken to transfer data from the memory unit to the computing unit, and the second is the data access time within the memory structure. It is tough to improve the data access time within the memory unit as the read and the write time of a memory technology are almost fixed due to constraints of physics [3]. Considering the DRAM technology, the capacity has increased nearly 128 times over the last two decades, which is now saturating. The BW has improved by 20-fold, while latency improvement is only 1.3 times [2]. The other issue, i.e., the latency due to data movement between the host processor and the memory, wastes nearly 60% of energy on-chip [4], so reducing data transfers or increasing the BW can give better results and is being targeted by the industry.

The motivation behind developing the 3D memory structures is to bridge the gap between processor BW requirements and available BW with conventional DRAM solutions. Using packaging technology of Through Silicon Via (TSV), memory structures like Hybrid Memory Cube (HMC) by Micron [5] and High Bandwidth Memory (HBM) [6] by Samsung, AMD, and SK Hynix have attempted to bridge that gap. The TSVs connecting the stacked DRAM layers provide a BW of nearly 320Gbps within the stacked memory structure. However, a comprehensive performance evaluation of HBM and a comparative study of HBM against GDDR6 technology is needed.

The paper aims to quantify the performance improvement observed in HBM over and above the GDDR5 and GDDR6 technologies. The performance improvement is discussed in terms of latency improvement and energy efficiency. The evaluation is performed for both synthetic benchmarks and high-performance computing (HPC) workloads.

The rest of the paper is organized as follows. Background

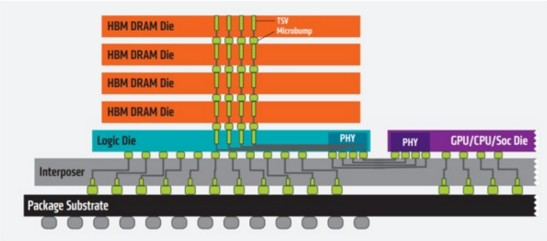


Fig. 2. High Bandwidth Memory [6]

details about HBM, GDDR5, and GDDR6 are presented in section II. The section III discusses brief literature on the characterization and evaluation of the latest 3D memory technologies. In section IV, we present the methodology used for our work. Section V has two parts. In part V-A we compare the performance between HBM and GDDR6 for synthetically generated traces. Section V-B examines the power and performance of real-world benchmarks. Finally, section VI shares brief conclusion and future prospects.

II. Background

This section presents a brief discussion on the GDDR5 technology and its successor technologies, HBM and GDDR6. However, GDDR6 technology is more mature compared to the HBM memory technology. On the one hand, the GDDR6 technology uses increased per pin BW (16Gbps) and 16-N prefetch for higher performance [7]. On the other hand, the HBM relies on the high BW available from the TSVs and the Silicon Interposers [6]. Due to the stacked structure, the HBM uses considerably lower power compared to GDDR6. Fig. 1 and fig. 2 presents an architectural diagram of the GDDR6 and HBM memory technologies respectively. A detailed discussion on the HBM and GDDR technologies is presented as follows.

A. High Bandwidth Memory

High Bandwidth Memory (HBM) is a JEDEC Standard characterized DRAM. As shown in fig. 2, the HBM consists of a base logic die at the bottom with stacked DRAM dies interconnected by TSVs. Using the stacked architecture, TSV, micro-bump, and 2.5D package technologies using silicon interposer HBM provide higher capacity, BW, and power efficiency than standard DRAMs [8].

HBM was proposed as a replacement to GDDR5 due to the off-chip I/O pin and power constraints of GDDR5. Current generation HBM which is HBM2 [6], has multiple 8 Gb core DRAM dies and eight channels of 128 I/Os as compared to 32 I/Os of GDDR5. Depending on the density of HBM (which could be 2, 4, or 8 GB), the number of stacks, the number of memory banks, and the channels in each stack are decided. HBM2 exhibits an external BW as high as 256 GB/s. Also, HBM introduces the concept of the pseudo channel, which divides one channel into two pseudo channels enabling the access of different rows and columns at a time.

B. GDDR5 SGRAM & GDDR6 SGRAM:

The GDDR SGRAM stands for Graphics Double Data Rate Synchronous Graphics Random Access Memory. GDDR5 and its successor GDDR6 [7], [9]are technologically mature main-memory technologies currently available for commercial purposes. These memories are recommended for high memory BW applications, such as high-performance computation, gaming, and workstations. The bit transfer rate of GDDR6 is 16Gb/s which is twice that of the 8Gb/s bit transfer rate of the GDDR5 technology. The GDDR5 technology uses the 8-N prefetch and DDR interface technology to achieve high performance [9]. GDDR6 uses 16-N prefetch as an upgrade to GDDR5 performance [7]. Both GDDR5 and GDDR6 have 32B granularity. However, in the case of GDDR6 technology, we have two channels, which allow for performance enhancement and minimal effort needed to perform technology transition [7], [9]. At the 21nm process node, GDDR6 is expected to consume more than 10 % less power than GDDR5 technology. GDDR6 technology incorporates all the power-saving schemes available in the GDDR5. Adding to it, the low power features of GDDR6 make it more energy efficient.

III. Related Works

This section presents a brief review of the literature related to the characterization of HBM memory technology. Some of the literature discussing the architectural advantages of 3D memory are as follows. G Loh *et al*. identified a) the benefits of three-dimensional stacking, b) the high memory BW which could be achieved using TSVs, and c) the overall speedup in performance because of multiple parallel accesses [10]. J Jeddeloh *et al*. [11] discussed the internal architecture of 3D memory and how it could improve the overall memory access latency. Li *et al*. [12] focuses on all the existing memory technologies and how the inherent memory level parallelism can reduce the memory access latency.

Jun *et al*. [8] and Lee *et al*. [13] have studied the benefits of the 3D stacked memory structure like HBM. The papers highlight the individual features of the 3D structure, such as the effect of stacking the layers, the performance of TSVs, and the two operating modes- the legacy mode and the pseudo channel mode. Li *et al*. [14] conducted a performance and energy-based exploration of the HBM memory technology. Asifuzzaman *et al*. [15] presented a timing analysis of the HBM and compared the it with the DDR technology. An experimental study for real HBM under reduced-voltage conditions has been performed by Larimi *et al*. in [16]. The experiment also gathers information on the HBM reliability when subjected to lower voltage conditions. Congiu *et al*. [17] performed fine-grained evaluation of the HBM for MPI library, as the problem focuses on the capacity constraint of the HBM. Runbin *et al*. focus on the data analytics workloads as they deploy the workload on FPGAs having HBM support [18].

The HPC workloads extensively utilize the main memory, as it operates on enormous data sets and often does not rely on the locality of data. Hence, excessive main memory accesses are observed. Although much effort has been given to characterize the HBM, not much effort in the literature is given to evaluating the performance of HBM with HPC workloads.

TABLE I
CPU CONFIGURATION

CPU	8 cores, Out-of-Order, 2.667GHz
L1I	Private, 32KB, 2-way assoc
L1D	Private, 64KB, 2-way assoc
L2	2MB, 8-way assoc
L3	16MB, 16-way assoc

TABLE II
SPECIFICATIONS OF GDDR5, GDDR6, AND HBM2

	GDDR5	GDDR6	HBM
Bankgroups	4	4	4
Banks per group	4	4	4
Device Width (bits)	32	16	128
Capacity (GB)	4	4	4
Number of Channels	1	1	8
Channel Size (MB)	4096	4096	512

This paper aims to bridge the gap in the literature by characterizing the performance of HBM technology against HPC workload and perform a comparative study against GDDR6 technology. We evaluate HPC workloads from RBS [19] and also perform the examination on synthetic benchmarks since the characteristics of synthetic benchmarks closely resemble the characteristics of HPC workloads [20].

IV. METHODOLOGY

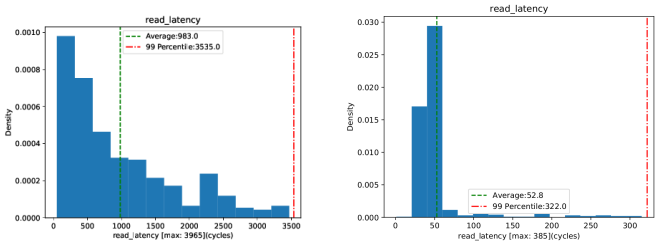
Our simulation setup consists of a Multicore CPU with x86 ISA, and consists of private L1(I & D) caches, partially shared L2 caches and shared L3 caches. The benchmarks under evaluation are placed on the CPU that connects to the memory under observation. The specifications of the CPU used are mentioned in table I. The specifications for GDDR5, GDDR6 and HBM2 under investigation are mentioned in table II.

We have examined the synthetic benchmarks only on DRAMSim3. The benchmarks were generated using a trace generator where random addresses were taken from a 4GB address space. However, for the High-Performance Computing Workloads from the RBS [19], we design our high-level architecture with the support of Structural Simulation Toolkit 11.0.0, to which we feed our data from the DRAMSim3 [21] simulator. We use the following two components - Ariel and MemHierarchy, from the Structural Simulation Toolkit 11.1.0 [22]. The workloads used from the RBS [19] are - CFD Solver, LavaMD, Back Propagation, Leukocyte, LU Decompostion, Myocyte, SRAD and Kmeans Algorithm.

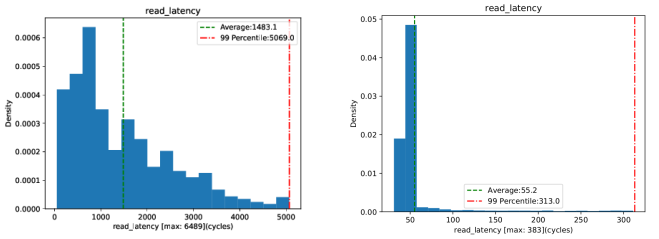
V. PERFORMANCE EVALUATION

This section presents a comparative study of HBM2 and GDDR6 in terms of performance and energy efficiency. In terms of BW, HBM2 far exceeds the conventional DRAMs. Sadagopan *et al*. in their work [23], presented that if system behavior is plotted as latency per request vs. actual BW usage (requests/unit time), three distinct regions are observed.

A similar behavior can be analyzed in the performance of the HBM2 and converntional DRAM technology. The latest version of the GDDR series, i.e. GDDR6 can provide high BW as HBM2 in less number of pins due to its very high

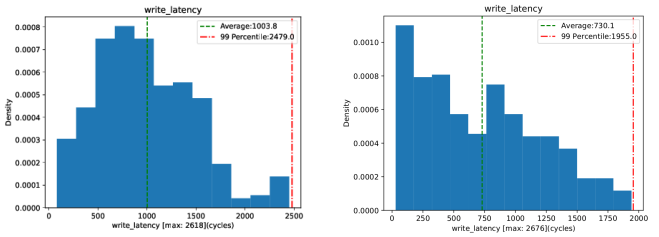


(a) GDDR6 : 500 instructions (b) HBM2 : 500 instructions

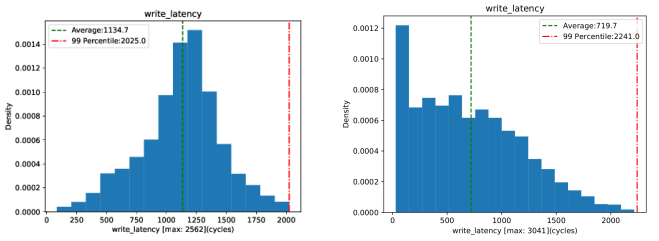


(c) GDDR6 : 5000 instructions (d) HBM2 : 5000 instructions

Fig. 3. Evaluation of GDDR5 and HBM across different number of read instructions



(a) GDDR6 : 500 instructions (b) HBM2 : 500 instructions



(c) GDDR6 : 5000 instructions (d) HBM2 : 5000 instructions

Fig. 4. Evaluation of GDDR6 and HBM across different number of write instructions

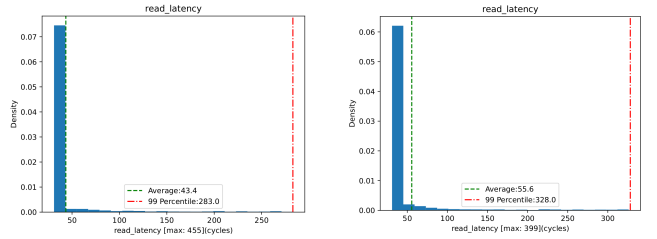
speed of 16Gb/s per pin; but it will require a number of such devices to match the desired BW due to the presence of only 32 pins in GDDR series. Comparing HBM2 with GDDR6 can give insight into the advantage of having 128-bit wide eight channels in the stacked memory structure.

A. Synthetic Benchmarks Test

A synthetic benchmark provides more control over the number of operations to perform within the given constraints. In the case of synthetic benchmarks, the traces generated do not tend to use any type of locality (either temporal or spatial). This behavior allows us to examine the memory system's capabilities to the fullest. We have fixed the number of cycles

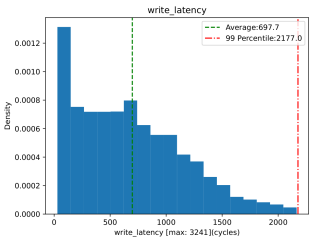
TABLE III
COMPARISON OF AVERAGE AND 99 PERCENTILE OBSERVATION OF READ LATENCY AND WRITE LATENCY FOR HBM2 AND GDDR6 TECHNOLOGIES

#Instructions	GDDR6				HBM2			
	Read Latency		Write Latency		Read Latency		Write Latency	
	Average	99 Percentile	Average	99 Percentile	Average	99 Percentile	Average	99 Percentile
500	983.0	3535	1003.8	2479	52.8	322	730.1	1955
1000	1342.3	3650	1160.6	2162	54.9	331	702	2055
2500	1410	3835	1090.2	2091	54.4	316	693.4	2191
5000	1483.1	5069	1134.7	2025	55.2	313	719.7	2241

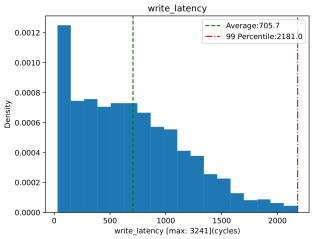


(a) HBM2 Read: Close Page

(b) HBM2 Read: Open Page

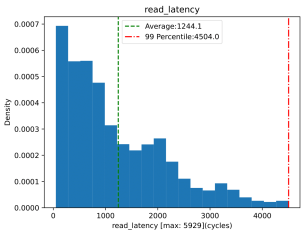


(c) HBM2 Write: Close Page

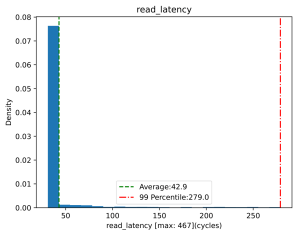


(d) HBM2 Write: Open Page

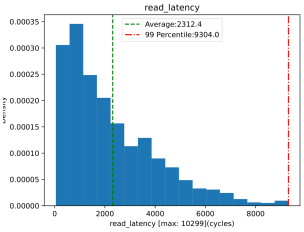
Fig. 5. Evaluation of HBM read and write latency for open and close page policy



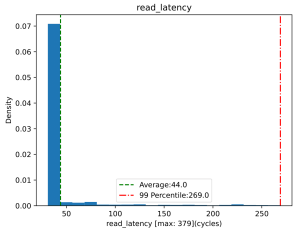
(a) GDDR6 : R/W : 2



(b) HBM2 : R/W : 2



(c) GDDR6 :R/W : 0.33



(d) HBM2 : R/W : 0.33

Fig. 6. Read latency evaluation of GDDR6 and HBM across different number of read to write ratio

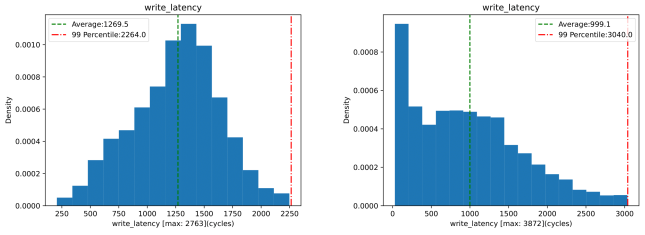
for all our simulations with synthetic benchmarks. In fig. 3, 4, 6 and 7, we have used a unique approach to highlight the performance of the HBM2 and the GDDR6 technology, where we plot the density of the number of read and write operations taking a certain number of cycles to execute. Then a statistical analysis is performed to identify the mean and 99th percentile of the read and write latencies.

In fig. 3 and 4, we present the read and write latency distributions respectively for HBM2 and GDDR6 memories. As stated already, we have fixed the number of cycles for which the simulations are performed, and then we have increased the number of read and write instructions. The orientation in performing such a simulation is to observe the system performance when put under increasing pressure. Table III presents added cases for 1000 and 2500 read and write instructions. We observe that the high BW of the TSVs (nearly 256GB/s) in the stacked memory outperforms the very high I/O pin BWs. The inherent network structure in the stacked memory provides the added benefits of memory level parallelism [20]. Hence, we observe that for 500 individual read and write instructions, the average read and write latency in GDDR6 memories is 983 cycles and 1003.8 cycles, respectively, whereas the average read and write latency in the case of HBM2 memory is only 55.2 cycles and 730.1 cycles respectively.

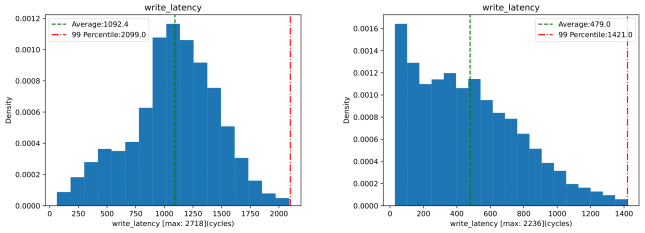
From the results in fig. 3 and table III, when the number of read instructions changed from 500 to 5000 (nearly tenfold), the average read latency for GDDR6 increased by 500 cycles. However, in the case of HBM2, the change is only 2.4 cycles. A similar pattern is observed in fig. 4 and table III for synthetically generated write operations. When the number of write instructions changes from 500 to 5000, the write latency changed by 130.9 cycles for GDDR6 while it was quite stable for HBM2 varying between 690 to 730 cycles, around 40 cycles. So even after a tenfold rise in load, HBM2 is still in a constant latency region for both read and write operations while the latency of GDDR6 has increased significantly. Increasing the number of instructions tenfold increases the request rate by the same factor, which stresses the available BW and latency increases significantly. The high BW because of the TSVs and pseudo channel mode of HBM2 helps it process more requests in a given unit of time, and hence the latency remains almost constant in varying loads.

In fig. 6 and fig. 7 the latency for read and write were noted for read to write ratio (R/W) = 0.33 and read to write ratio = 2 to see the impact of unequal distribution of read and write instructions. In both cases, HBM2 gave better latency compared to GDDR6. While GDDR6 does have a higher speed per pin, the amount of load it can handle in parallel is limited due to its low pin counts. In fig. 6, we observe that as the ratio of read to write increases from 0.33 to 2, the average read latency decrease from 2312.4 to 1244.1 cycles in GDDR6. On the other hand a very low decrease in the case of HBM2 from 44 to 42.9 cycles. An explanation for such a behavior is that a relatively less time needed to perform a read operation than the write operation. Note that a reliable write operation consumes more time and energy than a read operation. When observed statistically, the average read latency decreases when the density of read operations increases within the given simulation period. A similar conclusion can be derived from the behavior in fig. 7.

3D memory technologies often use smaller row buffers. The reason behind such a feature is to avoid the ”overfetch”



(a) GDDR6 : R/W : 2 (b) HBM2 : R/W : 2



(c) GDDR6: R/W : 0.33 (d) HBM2: R/W : 0.33

Fig. 7. Write latency evaluation of GDDR6 and HBM across different number of read to write ratio

problem [20]. The overfetch problem occurs when a complete DRAM row is brought to the row buffer, but only a few bits are used. For this reason, a closed page policy is preferred to further address the overfetch problem. Fig. 5 shows the read and write latency of HBM2 for open and close page policy. Synthetic benchmarks do not support any locality principle. Open page policy favors data locality as it repeatedly avoids unnecessary row activation for the same data accesses. However, without locality, this becomes a hindrance as the data not being present in the row buffer delays the activation and latency increases. In obtained results, close page HBM2 spends 43.4 and 697.7 cycles for read and write operations compared to 53.6 and 705.7 cycles in open page policy. Considering the synthetic benchmarks, the close page policy performs better than the open page policy as it avoids the delayed row activation due to compulsory misses.

B. Real Benchmark Test

Synthetic benchmarks give a fair idea about the performance gap between HBM2 and GDDR6 due to their BW difference and structural specifications. However, in the real world, the workloads do use locality principles quite frequently, and in that case, the performance will not be decided purely based on BW. Some real-world benchmarks like backprop, Kmeans, srad, cfd, lavaMD, lud, leukocyte, and myocyte were run on GDDR5, HBM2, and GDDR6, under the same capacity constraints as in the synthetic benchmarks investigation to get the overall performance characterization.

In fig. 8 performance of HBM2, GDDR5, and GDDR6 are observed for various real-world HPC benchmarks in terms of normalized execution time. In each benchmark, HBM2 performs better while the performance of GDDR5 and GDDR6 is almost at par, with GDDR6 being the faster technology that performs better in most benchmarks. Each benchmark has its own characteristics to exploit, and performance of

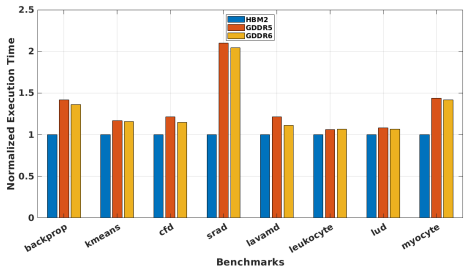


Fig. 8. Normalized execution time spent in executing different benchmarks with HBM2, GDDR5 and GDDR6

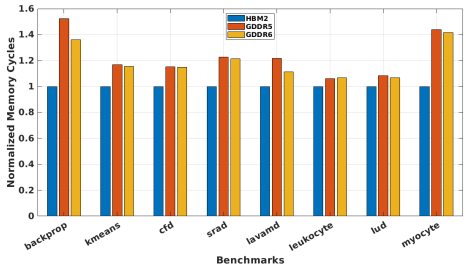


Fig. 9. Normalized memory cycles in executing different benchmarks with HBM2, GDDR5 and GDDR6

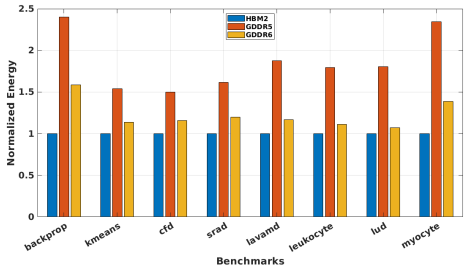
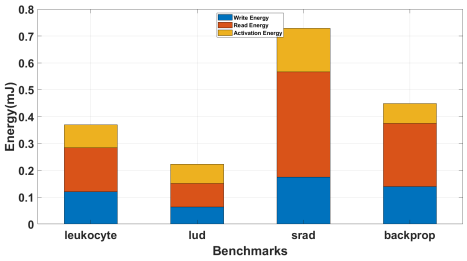
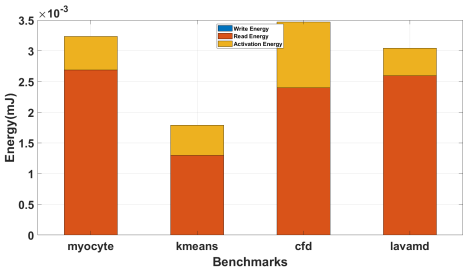


Fig. 10. Normalized energy spent for executing different benchmarks with HBM2, GDDR5 and GDDR6

each memory technology depends on that. Benchmarks like backprop and srad rely on really high data level parallelism [24] which HBM2 exploits with the high internal BW and the inherent internal network structure. Fig. 8 presents the normalized execution time needed to execute each benchmark, when subjected to different memory technologies. In backprop, HBM2 performs nearly 29.42% and 26.52% faster than GDDR5 and GDDR6, respectively. For srad, the improvement is 52.45% and 51.12%. The major takeaway from fig. 8 is the improvement in execution time across all benchmarks for HBM technology when compared to GDDR6, because of the high BW and parallelism support available in HBM memories. The normalized memory cycles are presented in fig. 9. Fig. 9 gives an idea regarding the internal BW, access speed, or latency comparison of each memory technology. HBM2, due to its 1024 TSVs [15] and 3D stacked structure accesses data extremely fast and gives really low latency; hence memory cycles spent is really low. In backprop, HBM2 spends nearly 34.38% less time in memory than GDDR5 and 23.47% compared to GDDR6. Although the improvements can



(a) HBM2 benchmarks: leukocyte, lud, srad, backprop



(b) HBM2 benchmarks: myocyte, kmeans, cfd, lavamd

Fig. 11. Evaluation of HBM2 read,write and activation energy distribution

vary depending on benchmarks, the HBM2 is always the better performer, followed by GDDR6 in most cases.

The major advantage of HBM2 is energy-saving which is shown in fig. 10. One of the key reasons for such huge energy savings is the relatively small chip footprint. HBM2 gives better energy efficiency for all the benchmarks than GDDR6, going as high as 28.13% in the myocyte benchmark. If compared with GDDR5, the highest improvement observed is 58.35% in backprop. GDDR6 gives better energy and execution performance than GDDR5. In energy, it saves nearly 33.9% and 41% compared to GDDR5 in backprop and myocyte, respectively. GDDR6 being the latest technology in GDDR series, is able to give more energy-efficient performance along with a high data rate per pin of nearly 16 Gb/s. The available data rate per pin in GDDR6 far exceeds the data rate per pin of HBM technology (2Gb/s). However, the I/O BW of GDDR6 is limited by the number of pins available, where on one hand GDDR6 has only 32 pins. On the other hand, the HBM2 has 128 pins per channel in eight channel configuration, thereby operating at 256GB/s [6], [7]. Despite the cumulative high BW avaliable in HBM2, the relatively low data rate per pin contributes to the energy efficiency.

The extent of data locality available in each benchmark can be interpreted from fig. 11, where the read, write, and activation energy spent for each benchmark is plotted. Read and write energy involves the energy spent on obtaining and modifying the data, respectively, in DRAM after accessing it. The energy spent to access the particular row or column where data is present is given as the activation energy. From our observation in fig. 11, the ratio between the energy spent on read & write and the energy spent on the activation gives a fair idea about the portion of time spent on the activation of a DRAM row to collect new data. This ratio is a direct indication of the data locality that the individual benchmark relies upon. Benchmarks involved in fig. 11(b) are trained data sets and meant for the read-only purpose, hence do not involve any write energy.

VI. Conclusions

In this paper, we quantify the performance improvement and energy efficiency of the HBM2 memory against more mature GDDR6 memory. Our investigation stresses the HBM system and observes that the read latency of HBM2 is nearly hundred times better than GDDR6. The write latency of HBM2 also performs better than the GDDR6 when subjected to the same workload. As we examine the HPC workloads from the RBS, we observe the 30% improvement in performance on average for HBM2 compared to GDDR5 and GDDR6 technology, with a maximum observable execution time improvement for SRAD benchmark with nearly 52%. HBM2 being the advanced technology, gives $\sim 23\%$ more energy-efficient performance on average across all benchmarks when compared to GDDR6.

References

[1] Wm A Wulf and Sally A McKee. Hitting the memory wall: Implications of the obvious. *ACM SIGARCH comp. arch. news*.
[2] J Henessey and D Patterson. Computer architecture: A quantitative approach mogran kaufman publishers. *Palo Alto, CA*, 1990.
[3] Neil HE Weste and David Harris. *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2015.
[4] Amirali Boroumand et al. Google workloads for consumer devices: Mitigating data movement bottlenecks. In *Proceedings of the 23rd International Conference on ASPLOS*, 2018.
[5] HMC Consortium et al. Hybrid memory cube specification 2.1. *Retrieved from hybridmemorycube. org*, 2013.
[6] Joe Macri. Amd's next generation gpu and high bandwidth memory architecture: Fury. In *2015 IEEE Hot Chips 27 Symposium (HCS)*, 2015.
[7] MICRON Technology. GDDR6 SGRAM. *Technical Note*, 2017.
[8] Hongshin Jun et al. Hbm (high bandwidth memory) dram technology and architecture. In *2017 IEEE International Memory Workshop (IMW)*.
[9] MICRON Technology. GDDR5 SGRAM. *Technical Note*, 2014.
[10] Gabriel H Loh. 3d-stacked memory architectures for multi-core processors. *ACM SIGARCH comp. arch. news*, 36(3).
[11] Joe Jeddeloh et al. Hybrid memory cube new dram architecture increases density and performance. In *2012 Symposium VLSIT*.
[12] Shang Li et al. A performance & power comparison of modern high-speed dram architectures. In *MEMSYS 2018*.
[13] Dong Uk Lee et al. Design considerations of hbm stacked dram and the memory architecture extension. In *2015 IEEE CICC*.
[14] Bingchao Li et al. Exploring new features of high-bandwidth memory for gpus. *IEICE Electronics Express*, 2016.
[15] Kazi Asifuzzaman et al. Demystifying the characteristics of high bandwidth memory for real-time systems. In *2021 IEEE/ACM ICCAD*.
[16] Seyed Saber Nabavi Larimi et al. Understanding power consumption and reliability of high-bandwidth memory with voltage underscaling. In *2021 DATE*.
[17] Giuseppe Congiu et al. Evaluating the impact of high-bandwidth memory on MPI communications. In *2018 IEEE 4th ICCC*, 2018.
[18] Runbin Shi et al. Exploiting hbm on fpgas for data processing. *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, 2021.
[19] Shuai Che et al. Rodinia: A benchmark suite for heterogeneous computing. In *2009 IEEE IISWC*, 2009.
[20] Paul Rosenfeld. *Performance exploration of the hybrid memory cube*. PhD thesis, 2014.
[21] Shang Li et al. Dramsim3: a cycle-accurate, thermal-capable dram simulator. *IEEE Computer Architecture Letters*, 19(2):106–109, 2020.
[22] Arun F Rodrigues et al. The structural simulation toolkit. *ACM SIGMETRICS Performance Evaluation Review*, 38(4):37–42, 2011.
[23] Sadagopan Srinivasan. *Prefetching vs The Memory System: Optimizations for Multi-Core Server Platforms*. PhD thesis, 2007.
[24] Leonardo Dagum et al. Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engg.*, 5(1).