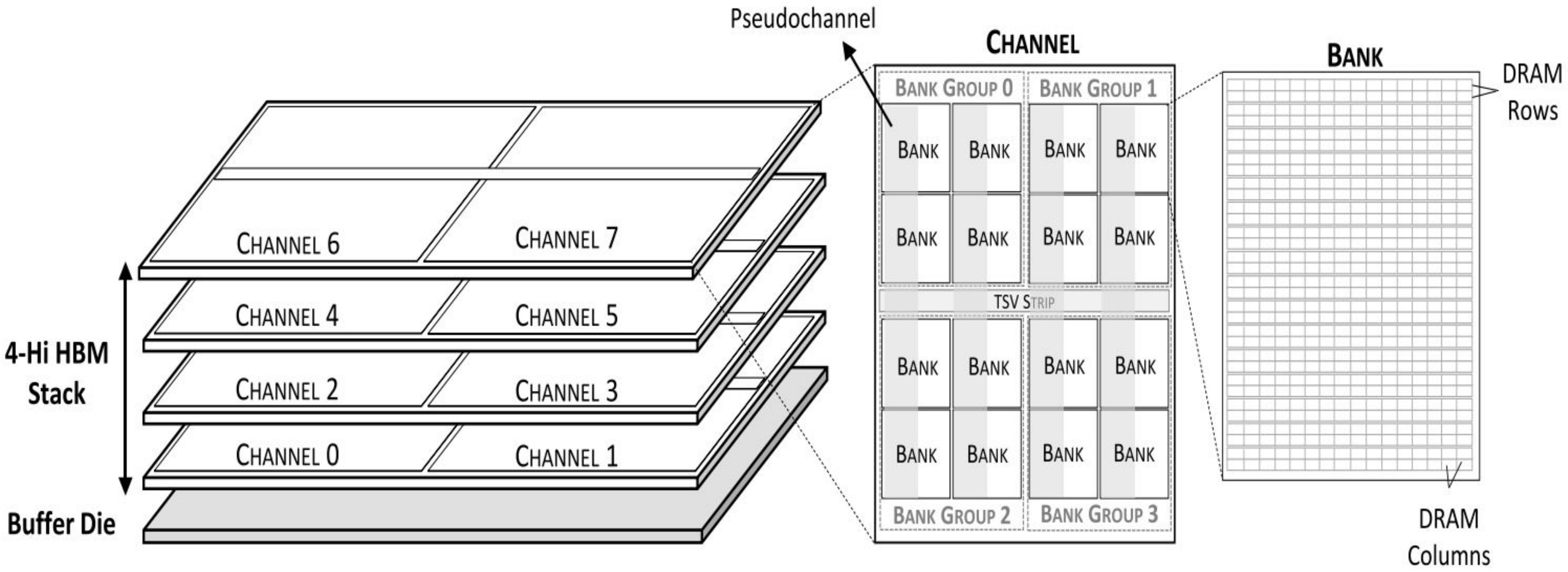
Thermal Mitigation for HBM Architecture

HBM2 Architecture



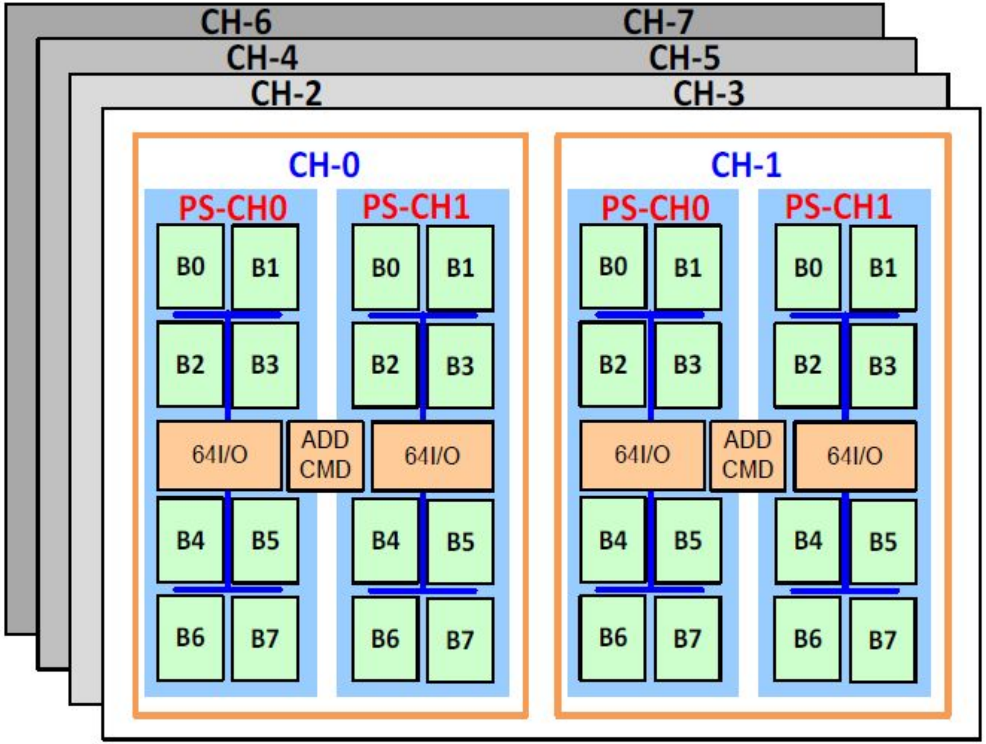
HBM Parameters From Previous Works

Parameter	**Value**
Number of layers/ DRAM dies	4
Number of channels	8 (2 per DRAM die)
Memory controller	1 per channel
Ranks, banks, bank groups	1 rank/channel 2 bank groups/rank 4 banks/bank group
Memory size	1 GB
Memory mapping	rorabgbachco

HBM2 Parameters From Previous Works

Parameter	**Value**
Number of layers/ DRAM dies	4
Number of channels	8 (2 per DRAM die)
Memory controller	1 per channel
Ranks, banks, bank groups	1 rank/channel 4 bank groups/rank 4 banks/bank group
Memory size	4 GB
Memory mapping	rorabgbachco

Pseudo-channel in HBM2



Generating DRAM Access Trace For HBM2

- To determine the bank accessed for every DRAM request, we will use below masks and change the equation corresponding to HMC architecture in current infrastructure
    - channel_mask = 7 [111]
    - rank_mask = 0 [0]
    - bankgroup_mask = 3 [11]
    - bank_mask = 3 [11]
    - row_mask = 16383 [11111111111111]
    - column_mask = 31 [11111]

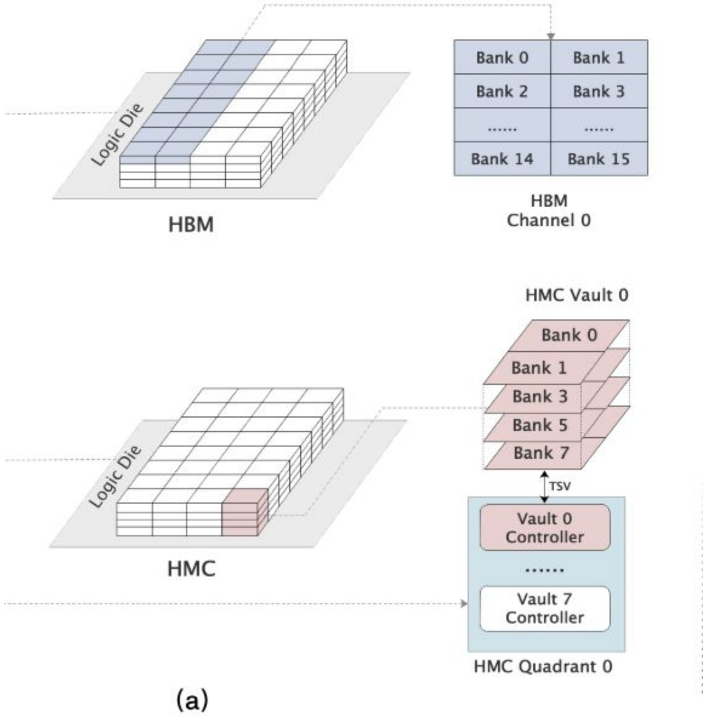
References Used For Determining HBM Parameters

- Shashank Adavally and Krishna Kavi. 2020. Towards Application-Specific Address Mapping for Emerging Memory Devices. *The International Symposium on Memory Systems*. Association for Computing Machinery, New York, NY, USA, 105–113. DOI:https://doi.org/10.1145/3422575.3422785
- Mahzabeen Islam, Soumik Banerjee, Mitesh Meswani, and Krishna Kavi. 2016. Prefetching as a Potentially Effective Technique for Hybrid Memory Optimization. In *Proceedings of the Second International Symposium on Memory Systems* (*MEMSYS '16*). Association for Computing Machinery, New York, NY, USA, 220–231. DOI:https://doi.org/10.1145/2989081.2989129
- Shang Li, Dhiraj Reddy, and Bruce Jacob. 2018. A performance & power comparison of modern high-speed DRAM architectures. In Proceedings of the International Symposium on Memory Systems (MEMSYS '18). Association for Computing Machinery, New York, NY, USA, 341–353. DOI:https://doi.org/10.1145/3240302.3240315
- X. Wang, A. Tumeo, J. D. Leidel, J. Li and Y. Chen, "HAM: Hotspot-Aware Manager for Improving Communications With 3D-Stacked Memory," in IEEE Transactions on Computers, vol. 70, no. 6, pp. 833-848, 1 June 2021, doi: 10.1109/TC.2021.3066982.
- S. Yin et al., "Parana: A Parallel Neural Architecture Considering Thermal Problem of 3D Stacked Memory," in IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 1, pp. 146-160, 1 Jan. 2019, doi: 10.1109/TPDS.2018.2858230.

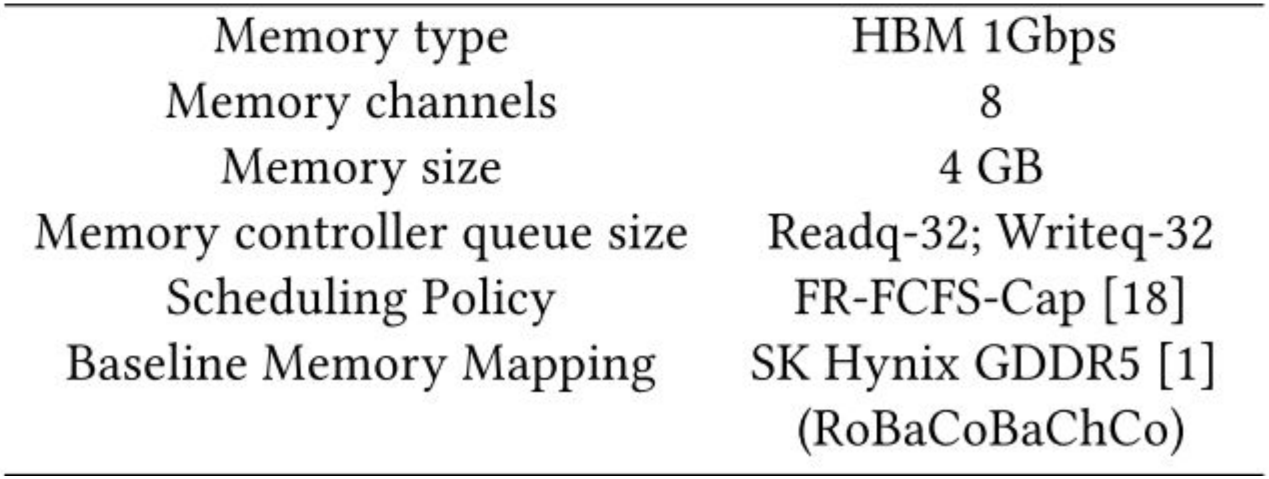
References Used For Determining HBM Parameters

- Mahzabeen Islam, Shashank Adavally, Marko Scrbak, and Krishna Kavi. 2020. On-the-fly Page Migration and Address Reconciliation for Heterogeneous Memory Systems. J. Emerg. Technol. Comput. Syst. 16, 1, Article 10 (February 2020), 27 pages. DOI:https://doi.org/10.1145/3364179
- Chiachen Chou, Aamer Jaleel, and Moinuddin Qureshi. 2017. BATMAN: techniques for maximizing system bandwidth of memory systems with stacked-DRAM. In Proceedings of the International Symposium on Memory Systems (MEMSYS '17). Association for Computing Machinery, New York, NY, USA, 268–280. DOI:https://doi.org/10.1145/3132402.3132404
- M. M. Rafique and Z. Zhu, "Memory-Side Prefetching Scheme Incorporating Dynamic Page Mode in 3D-Stacked DRAM," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 11, pp. 2734-2747, 1 Nov. 2021, doi: 10.1109/TPDS.2020.3044856.
- S. Adavally and K. Kavi, "3D-DRAM Performance for Different OpenMP Scheduling Techniques in Multicore Systems," *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018, pp. 675-683, doi: 10.1109/HPCC/SmartCity/DSS.2018.00119.

HBM vs HMC Architecture

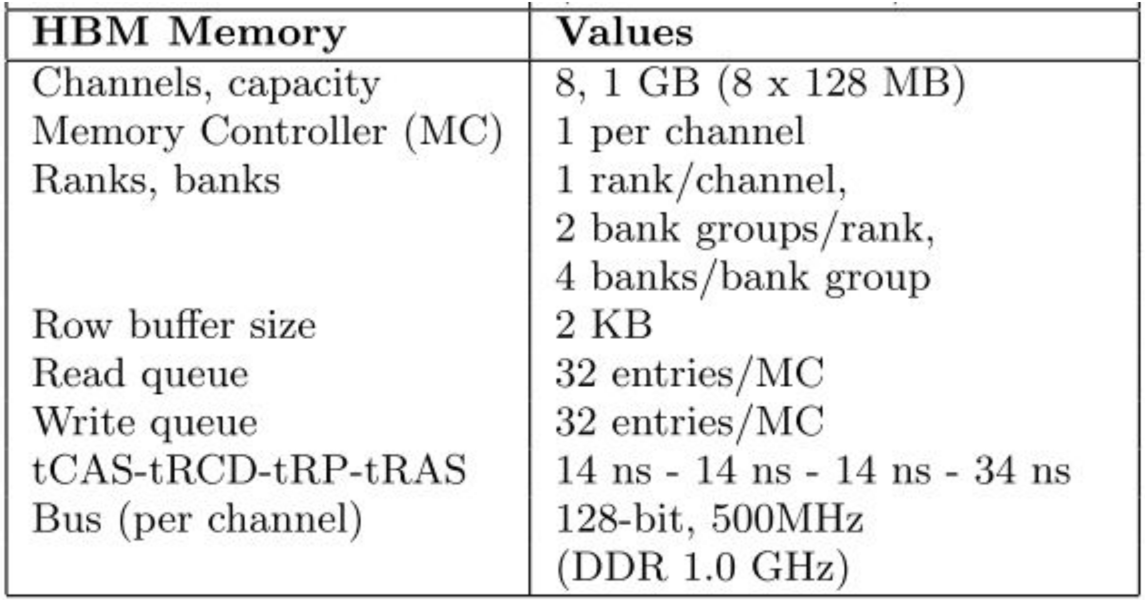


Towards Application-Specific Address Mapping for Emerging Memory Devices



Memory type	HBM 1Gbps
Memory channels	8
Memory size	4 GB
Memory controller queue size	Readq-32; Writeq-32
Scheduling Policy	FR-FCFS-Cap [18]
Baseline Memory Mapping	SK Hynix GDDR5 [1] (RoBaCoBaChCo)

Prefetching as a Potentially Effective Technique for Hybrid Memory Optimization



HBM Memory	Values
Channels, capacity	8, 1 GB (8 x 128 MB)
Memory Controller (MC)	1 per channel
Ranks, banks	1 rank/channel, 2 bank groups/rank, 4 banks/bank group
Row buffer size	2 KB
Read queue	32 entries/MC
Write queue	32 entries/MC
tCAS-tRCD-tRP-tRAS	14 ns - 14 ns - 14 ns - 34 ns
Bus (per channel)	128-bit, 500MHz (DDR 1.0 GHz)

A Performance & Power Comparison of Modern High-Speed DRAM Architectures

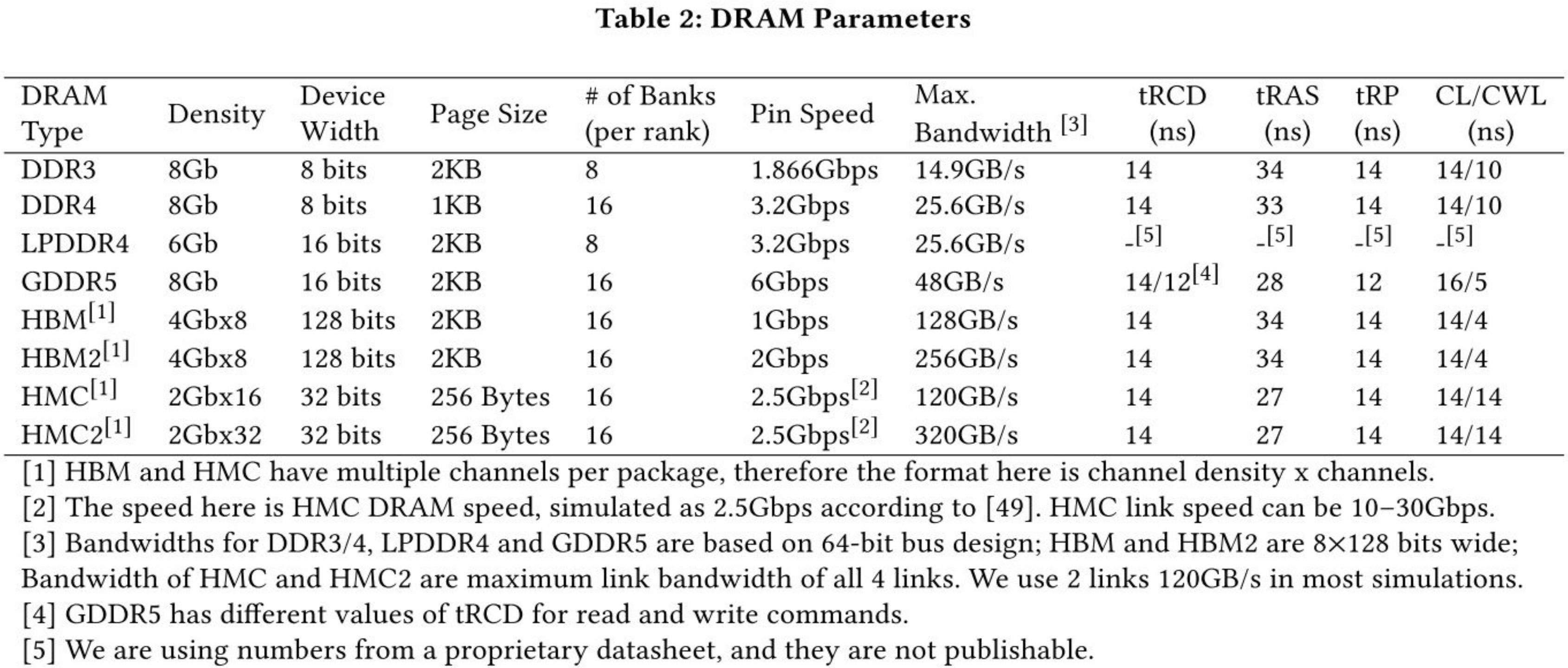


Table 2: DRAM Parameters

DRAM Type	Density	Device Width	Page Size	# of Banks (per rank)	Pin Speed	Max. Bandwidth [3]	tRCD (ns)	tRAS (ns)	tRP (ns)	CL/CWL (ns)
DDR3	8Gb	8 bits	2KB	8	1.866Gbps	14.9GB/s	14	34	14	14/10
DDR4	8Gb	8 bits	1KB	16	3.2Gbps	25.6GB/s	14	33	14	14/10
LPDDR4	6Gb	16 bits	2KB	8	3.2Gbps	25.6GB/s	-[5]	-[5]	-[5]	-[5]
GDDR5	8Gb	16 bits	2KB	16	6Gbps	48GB/s	14/12[4]	28	12	16/5
HBM[1]	4Gbx8	128 bits	2KB	16	1Gbps	128GB/s	14	34	14	14/4
HBM2[1]	4Gbx8	128 bits	2KB	16	2Gbps	256GB/s	14	34	14	14/4
HMC[1]	2Gbx16	32 bits	256 Bytes	16	2.5Gbps[2]	120GB/s	14	27	14	14/14
HMC2[1]	2Gbx32	32 bits	256 Bytes	16	2.5Gbps[2]	320GB/s	14	27	14	14/14

[1] HBM and HMC have multiple channels per package, therefore the format here is channel density x channels.

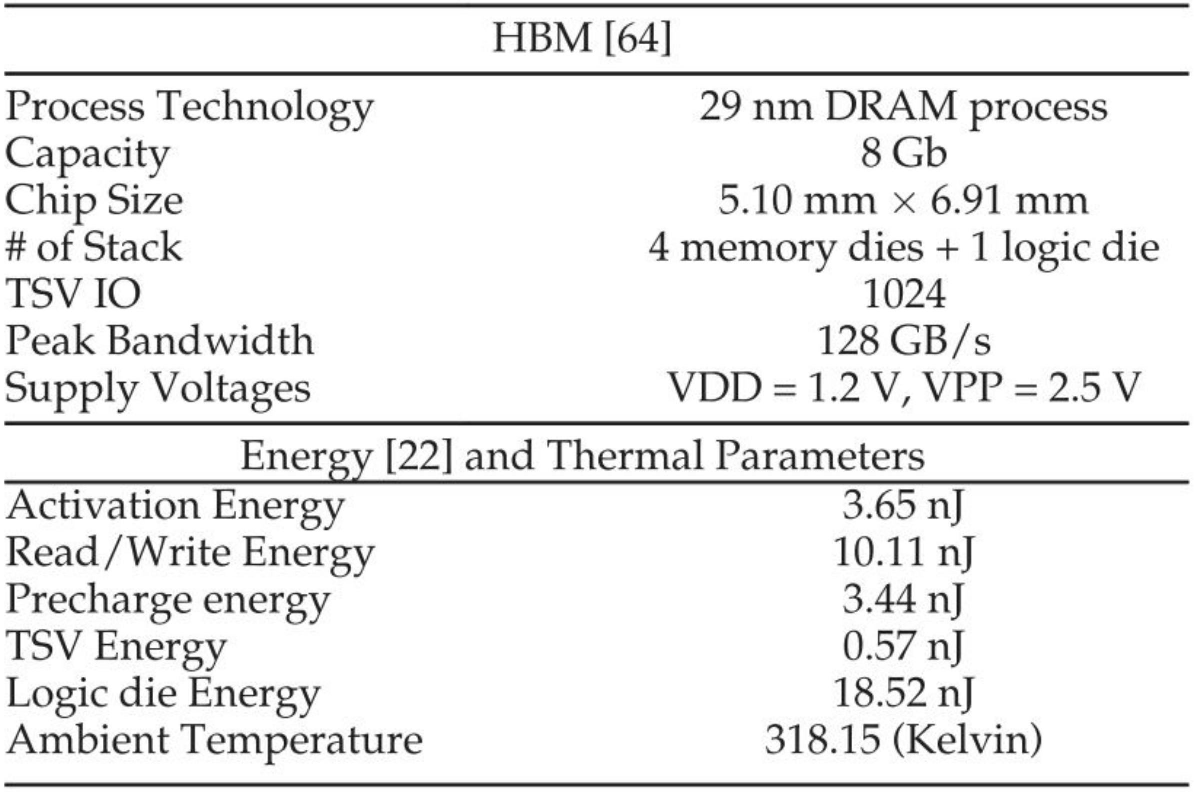
[2] The speed here is HMC DRAM speed, simulated as 2.5Gbps according to [49]. HMC link speed can be 10–30Gbps.

[3] Bandwidths for DDR3/4, LPDDR4 and GDDR5 are based on 64-bit bus design; HBM and HBM2 are 8×128 bits wide; Bandwidth of HMC and HMC2 are maximum link bandwidth of all 4 links. We use 2 links 120GB/s in most simulations.

[4] GDDR5 has different values of tRCD for read and write commands.

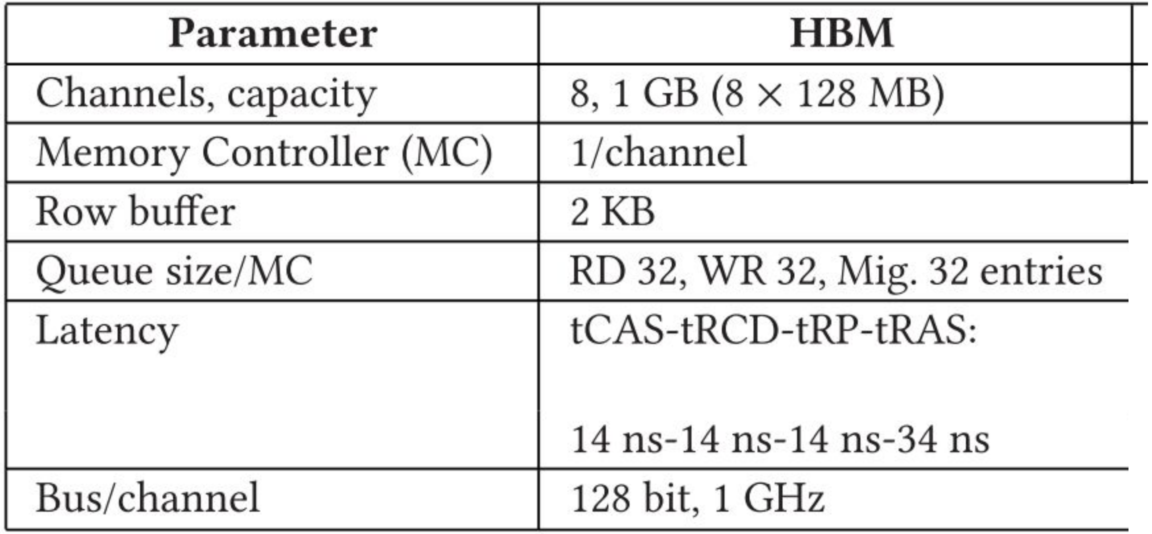
[5] We are using numbers from a proprietary datasheet, and they are not publishable.

Parana: A Parallel Neural Architecture Considering Thermal Problem of 3D Stacked Memory



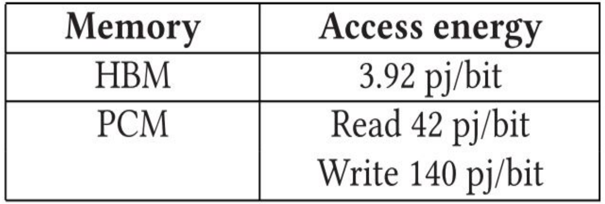
HBM [64]	
Process Technology	29 nm DRAM process
Capacity	8 Gb
Chip Size	5.10 mm × 6.91 mm
# of Stack	4 memory dies + 1 logic die
TSV IO	1024
Peak Bandwidth	128 GB/s
Supply Voltages	VDD = 1.2 V, VPP = 2.5 V
Energy [22] and Thermal Parameters	
Activation Energy	3.65 nJ
Read/Write Energy	10.11 nJ
Precharge energy	3.44 nJ
TSV Energy	0.57 nJ
Logic die Energy	18.52 nJ
Ambient Temperature	318.15 (Kelvin)

On-the-fly Page Migration and Address Reconciliation for Heterogeneous Memory Systems



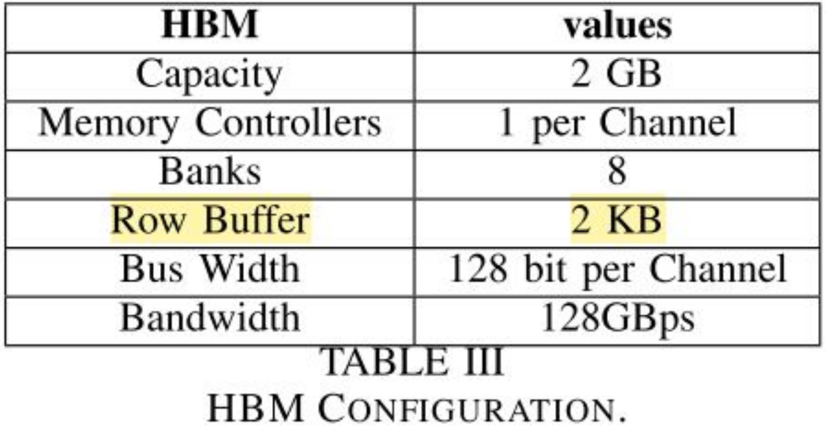
Parameter	HBM
Channels, capacity	8, 1 GB (8 × 128 MB)
Memory Controller (MC)	1/channel
Row buffer	2 KB
Queue size/MC	RD 32, WR 32, Mig. 32 entries
Latency	tCAS-tRCD-tRP-tRAS: 14 ns-14 ns-14 ns-34 ns
Bus/channel	128 bit, 1 GHz

Table 5. Memory Energy Parameters



Memory	Access energy
HBM	3.92 pj/bit
PCM	Read 42 pj/bit Write 140 pj/bit

3D-DRAM Performance for Different OpenMP Scheduling Techniques in Multicore Systems



HBM	values
Capacity	2 GB
Memory Controllers	1 per Channel
Banks	8
Row Buffer	2 KB
Bus Width	128 bit per Channel
Bandwidth	128GBps

TABLE III
HBM CONFIGURATION.

Important Parameters

- Energy_per_access
- Energy_per_refresh_access
- Bank_size [provided]
- No_columns
- No_bits_per_column
- T_refi
- No_refesh_commands_in_t_refw
- Leakage power equation

To-Do

- ISSC HBM papers
  - Die photograph
  -