TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

André Berthold

Fakultät Informatik, Computer Science, Institut für Systemarchitektur, Professur für Betreibssystems

Heterogeneous Memory Systems

Distributed Operating Systems // Dresden, June 23, 2025

Outline

Non-functional Properties

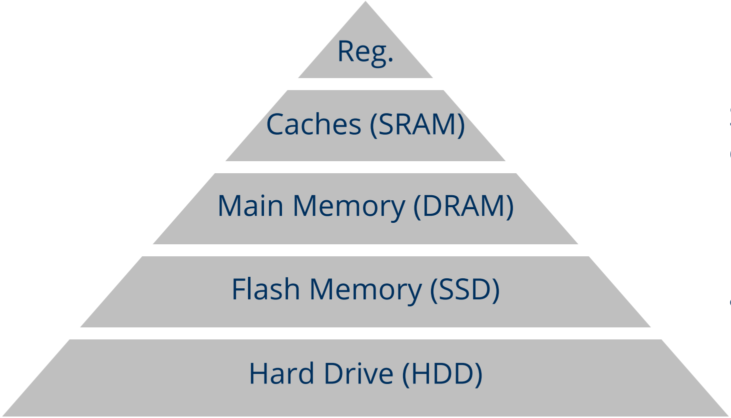
Add Heterogeneity

- NUMA (Recap)
- CXL
- HBM
- NVM
- RTM
- ...

Heterogeneous Memory Systems

Non-functional Properties

The Memory Hierarchy



So what is the difference between them?

In the end, SRAM, DRAM, SSDs, …, they are all memory.

Functional and Non-Functional Properties

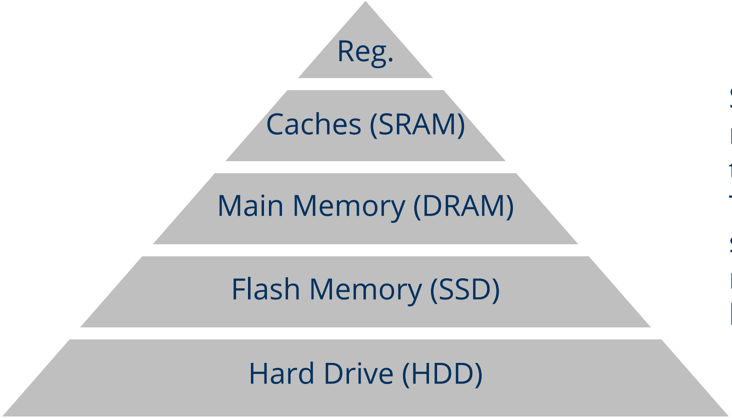
Functional property of a memory
memorize(/store) data

So from a functional point of view:

intel OPTANE [1] = just another ordinary main-memory technology

Non-Functional properties
• Capacity • Throughput • Latency • Persistence • Fault Tolerance • Wearout • Power consumption • …

The Memory Hierarchy

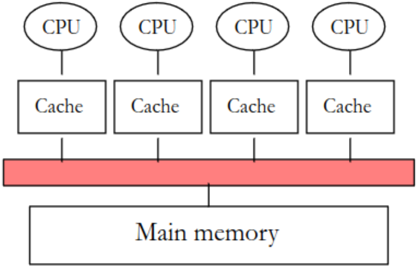


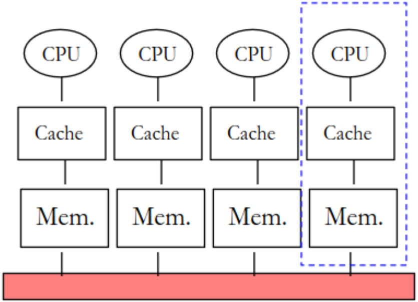
So they differ in their non-functional properties (Capacity, Latency, Troughput), which strictly increase as we move up or down the hierarchy.

Add Heterogeneity

Non-Uniform Memory Access

UMA versus NUMA





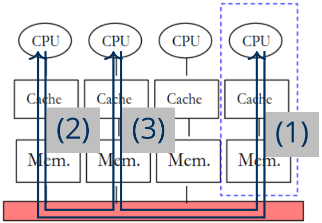
[2]

NUMA - Cache Coherence

NUMA types

- cache-coherent NUMA (ccNUMA)
  - guarantees system-wide cache coherence
  - uses directory-based cache-coherence protocol
- non-cache-coherent NUMA (ncNUMA)
  - no cache coherence guarantees
  - local memory access: goes through cache
  - remote memory access: by-passes cache

Directory-Based Cache-Coherence



[2]

(1) CPU4 request data from CPU1
   - CPU1 looks up in cache directory

(2) CPU1 forwards request to CPU2, which holds the cache line

(3) CPU2 answers CPU4

NUMA - Behavior

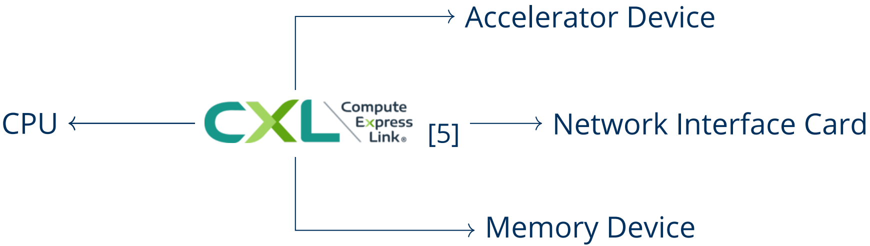
SGI UV 2000
• 64 Sockets
• 512 Cores

[3] [3]

Compute Express Link

"Compute Express Link (CXL)

is the broadly supported industry standard solution that has been developed to provide **low-latency**, **memory cache coherent** links between **processors**, **accelerators** and **memory devices**."   [4]



Compute Express Link - Protocols and Devices

CXL builds upon the physical interface of PCIe (5.0) and implements three protocols to replace the PCIe protocol:

- **CXL.io:** functionally equivalent to PCIe protocol; adopts to the interface
- **CXL.cache:** enables devices to (efficiently) access and cache host memory
- **CXL.memory:** enables host to access device attached memory

In CXL distinct three different device types:

Type 1	Type 2	Type 3
Accelerators without local memory	GPUs, ASICs, FPGAs with local memory	Memory devices providing additional memory space

Compute Express Link - Communication

Plugin

On device plugin (as CXL builds upon the physical interface of PCIe) both device negotiate about used protocols' via PCIe 1.0 (2.5 GT/s). They use the highest CXL/PCIe version both support.

Version	Speed	PCIe phys. layer	Lanes
CXL 1.1	**32 GT/s(64 GB/s)**	5.0	16-lane link
CXL 2.0	32 GT/s(64 GB/s)	5.0	16-lane link
CXL 3.1	**64 GT/s(128 GB/s)**	6.1	16-lane link

CXL - Use Case - Memory Disaggregation

Disaggregated Memory is a prominent and obvious use case for CXL.:

- D. Gouk et al., “Memory Pooling With CXL” (2023) [6]
- M. K. Aguilera et al., “Memory disaggregation: why now and what are the challenges” (2023) [7]
- A. Geyer et al., “Working with Disaggregated Systems. What are the Challenges and Opportunities of RDMA and CXL?” (2023) [8]

Rcmp: Reconstructing RDMA-based Memory Disaggregation via CXL [9]

RDMA	CXL
> 10 GiB/s	≤ 128 GiB/s
2 μs	250 ns

but

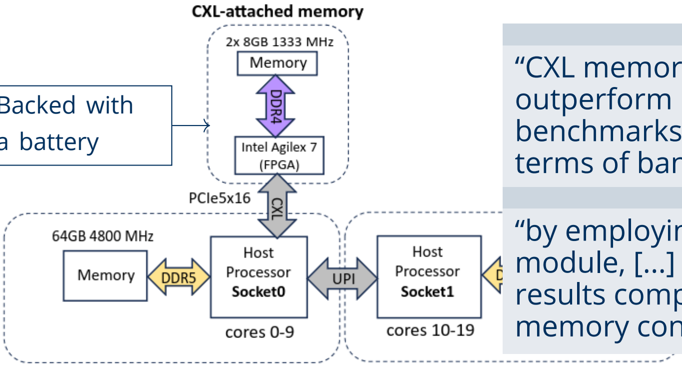
“[...]existing CXL-based approaches have physical distance limitation and cannot be deployed across racks.” [9]

“Rcmp [...] leverages RDMA to overcome CXL’s distance limitation.” [9]

CXL - Use Case - Persistent Memory

CXL Memory as Persistent Memory for Disaggregated HPC: A Practical Approach [10]

Backed with a battery

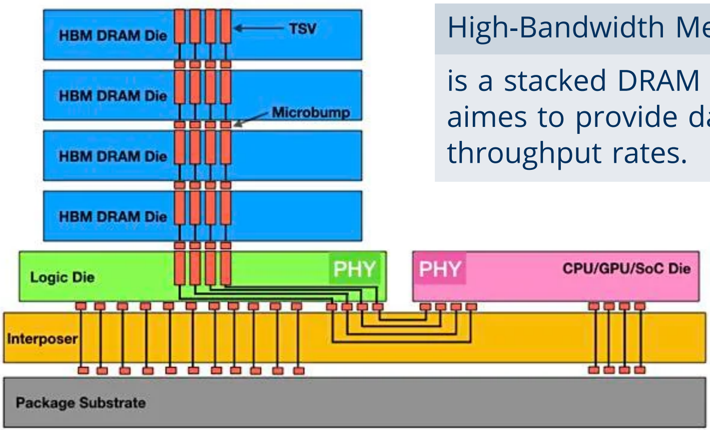


[10]

"CXL memory has the capability to outperform previously published benchmarks for Optane DCPMM in terms of bandwidth" [10]

"by employing a CXL-DDR4 memory module, [...] we achieved bandwidth results comparable to local DDR4 memory configurations" [10]

High-Bandwidth Memory

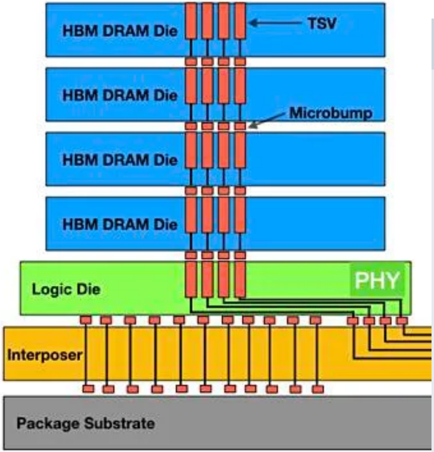


High-Bandwidth Memory (HBM)

is a stacked DRAM technology that aimes to provide data at high throughput rates.

[11]

High-Bandwidth Memory - Building Parts

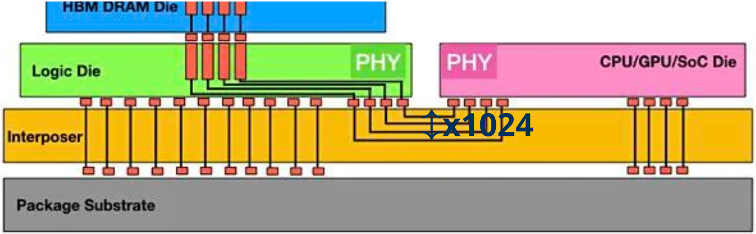


Building Parts

- **HBM-DRAM dies**
- **Through silicon via's (TSVs):** Electronic connection that extends through the silicon base of the dies
- **Microbumps:** Connection between two dies
- **Logic die:** Implements HBM-controll logic

[11]

High-Bandwidth Memory - Theoretical Throughput

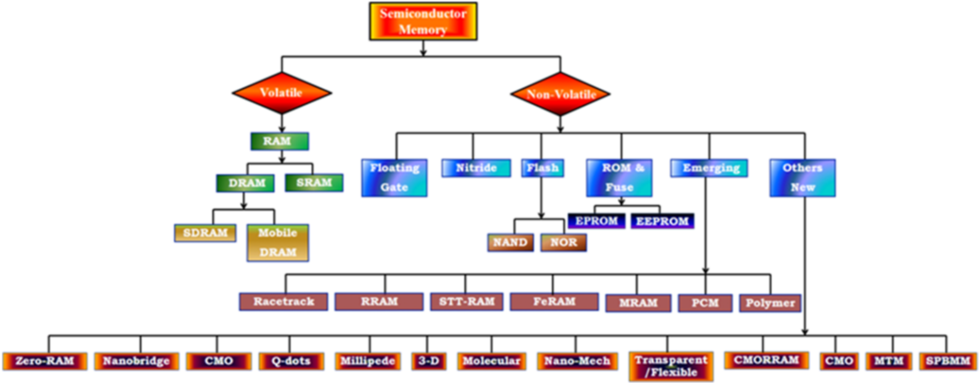


[11] [12], [13]

	HBM	HBM2	HBM2e	HBM3	DDR5
Bus clocking	1 GHz	2 GHz	3.6 GHz	6.4 GHz	4.8 GHz
Bus width	1024 bit	1024 bit	1024 bit	1024 bit	64 bit
Bandwidth	128 GB/s	256 GB/s	460.8 GB/s	819.2 GB/s	38.4 GB/s

Non-Volatile (Random-Access) Memory

There is a number of non-volatile memory technologies



[14]

Phase-Change Memory

The concept of phase-changing material storing information is relatively old.

 [15]

CDs, (later also DVDs, and Blu-rays)

Read CDs

Measuring reflection of a laser from the data carrying material.

Write CDs

Change the optical properties of the data carrying material by heating it with a laser.

Phase-Change Memory - Building Parts

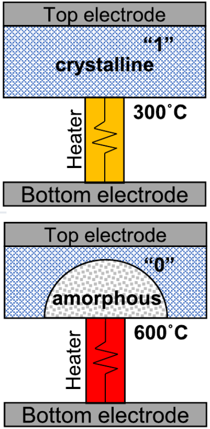
PCM does the reading and writing electronically rather than optically.

PCM cell building parts

- Bottom and top electrode
- Heater element
- Layer of phase-changing material (usually $Ge_2Sb_2Te_5$, same as in CD-RW, or DVD-RW [14])

The data is encoded in the material, which either is in

- a crystalline state (→ low resistance, 1) or
- an amorphous state (→ high resistance, 0) [16].

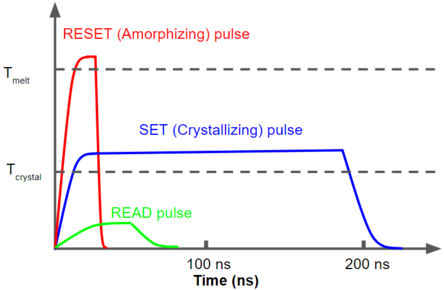


[16]

Phase-Change Memory - Data Access

Read

Short, low energy pulse to measure the resistance of the material.



[17]

Write

Set
Long, medium energy pulse to allow the material to organize in a crystalline pattern.

Reset
Very short, high energy pulse to melt the material. Cool down rapidly to avoid reorganization in a crystalline pattern. → amorphous state [17].

Phase-Change Memory - Bridging the Gap between Memory and Storage?

	DRAM	PCM	NAND-Flash
Cell Size (F^2)	$4 - 8$	$4 - 12$	$1 - 5$
Write endurance	$\geq 10^{16}$	10^{9}	$\leq 10^{5}$
Read time (ns)	30	50	25×10^{3}
Write time (ns)	50	$30 - 200$	$10^{5} - 10^{6}$

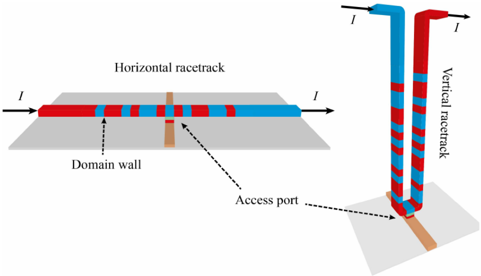
Adapted from: [18]

Race Track Memory - Building Parts

RTM Building Parts

- magnetic nanowires organized
  - horizontally or
  - vertically
- access ports that read and write data

Bits are stored on the nanowire as magnetization pointing up (0) or down (1).
→ allows unprecedented density [18]



[18]

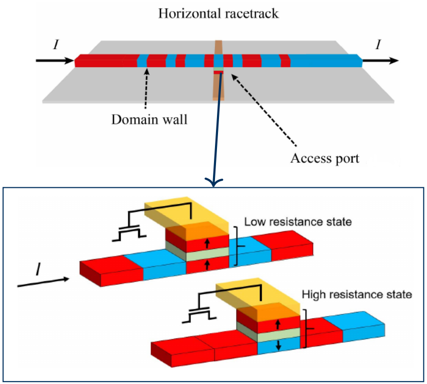
Race Track Memory - Data Access I

An access port typically consist of

- an access transistor (dominates size) and
- a magnetic tunnel junction (MTJ) made up of:
  - fixed orientation magnetic layer
  - insulating layer (typ. MgO or AL_2O_3)
  - section of the magnetic nanowire

Read out by magneto-resistive effects e.g.:

- Giant magnetoresistance (GMR)
- Tunneling magnetoresistance (TMR)
- **Tunneling magnetoresistance (TMR)** → up to 6-times GMT at room temperature [18]



[18]

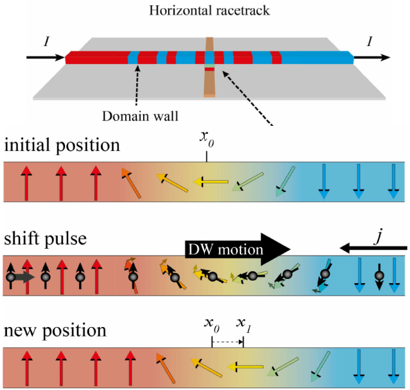
Race Track Memory - Data Access II

Write with larger currents that reorient the magnetic bit in the access port via spin transfer torques (STTs)

Shift

- Magnetic domain walls are shifted by a current pulse
- This rotates the local magnetization
- Magnetization via STT → electrons transfer their angular momentum to the localized magnetic momentum[a] [18]

[a] Applies for RTM versions 1.0 and 2.0



[18]

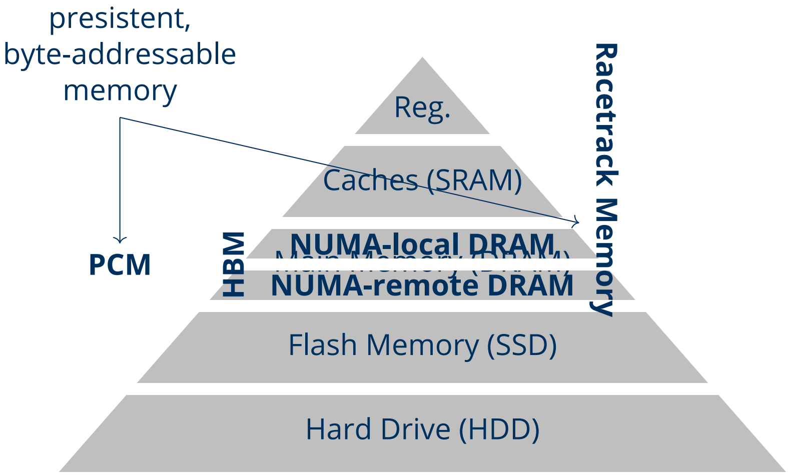
Phase-Change Memory - Comparison

	SRAM	RTM	DRAM	PCM	NAND-Flash
Cell Size (F^2)	120 - 200	≤ 2	$4 - 8$	$4 - 12$	$1 - 5$
Write endurance	$\geq 10^{16}$	$\geq 10^{16}$	$\geq 10^{16}$	10^{9}	$\leq 10^{5}$
Read time (ns)	$1 - 100$	$3 - 250$	30	50	25×10^{3}
Write time (ns)	$1 - 100$	$3 - 250$	50	$30 - 200$	$10^{5} - 10^{6}$

Adapted from: [18]

Heterogeneous Memory Systems

The New Memory ~~Hierarchy~~ Landscape



References I

[1] Intel, "Intel xeon max cpu is the sapphire rapids hbm line", Intel. (), [Online]. Available: `https://www.intel.de/content/www/de/de/products/docs/memory-storage/solid-state-drives/data-center-ssds/optane-ssd-p5800x-p5801x-brief.html` (visited on 07/14/2024).

[2] D. V. Nagarajan, *Lect. 4: Shared memory multiprocessors*, 2017. [Online]. Available: `https://www.inf.ed.ac.uk/teaching/courses/pa/Notes/lecture04-multi.pdf`.

References II

[3] W. Lehner, “Modern analytical database technology”, TU Dresden. (2014), [Online]. Available: https://de.slideshare.net/slideshow/wolfgang-lehner-technische-universitat-dresden/41254088#8 (visited on 07/13/2024).

[4] R. Press, “Compute express link (cxl): All you need to know”, (2024), [Online]. Available: https://www.rambus.com/blogs/compute-express-link/.

[5] computeexpresslink, “Cxl-homepage”, computeexpresslink.org. (), [Online]. Available: https://computeexpresslink.org/ (visited on 07/14/2024).

References III

[6] D. Gouk, M. Kwon, H. Bae, S. Lee, and M. Jung, "Memory pooling with cxl", *IEEE Micro*, vol. 43, no. 2, pp. 48–57, 2023. DOI: 10.1109/MM.2023.3237491.

[7] M. K. Aguilera, E. Amaro, N. Amit, E. Hunhoff, A. Yelam, and G. Zellweger, "Memory disaggregation: Why now and what are the challenges", *SIGOPS Oper. Syst. Rev.*, vol. 57, no. 1, pp. 38–46, Jun. 2023, ISSN: 0163-5980. DOI: 10.1145/3606557.3606563. [Online]. Available: https://doi.org/10.1145/3606557.3606563.

[8] A. Geyer, D. Ritter, D. H. Lee, *et al.*, "Working with disaggregated systems. what are the challenges and opportunities of rdma and cxl?", in *BTW 2023*, Bonn: Gesellschaft für Informatik e.V., 2023, pp. 751–755, ISBN: 978-3-88579-725-8. DOI: 10.18420/BTW2023-47.

References IV

[9] Z. Wang, Y. Guo, K. Lu, *et al.*, "Rcmp: Reconstructing rdma-based memory disaggregation via cxl", *ACM Trans. Archit. Code Optim.*, vol. 21, no. 1, Jan. 2024, ISSN: 1544-3566. DOI: 10.1145/3634916. [Online]. Available: https://doi.org/10.1145/3634916.

[10] Y. Fridman, S. Mutalik Desai, N. Singh, T. Willhalm, and G. Oren, "Cxl memory as persistent memory for disaggregated hpc: A practical approach", in *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, ser. SC-W '23, Denver, CO, USA: Association for Computing Machinery, 2023, pp. 983–994, ISBN: 9798400707858. DOI: 10.1145/3624062.3624175. [Online]. Available: https://doi.org/10.1145/3624062.3624175.

References V

[11] C. Mellor, “Dram, it stacks up: Sk hynix rolls out 819gb/s hbm3 tech”, (2021), [Online]. Available: https://www.theregister.com/2021/10/20/sk_hynix_hbm3/.

[12] R. Smith, “Sk hynix announces its first hbm3 memory: 24gb stacks, clocked at up to 6.4gbps”, (2021), [Online]. Available: https://www.anandtech.com/show/17022/sk-hynix-announces-its-first-hbm3-memory-24gb-stacks-at-up-to-64gbps.

References VI

[13] K. Joonyoung and K. Younsu, “Hbm: Memory solution for bandwidth-hungry processors”, SK hynix Inc. (2014), [Online]. Available: https://web.archive.org/web/20150424141343/http://www.setphaserstostun.org/hc26/HC26-11-day1-epub/HC26.11-3-Technology-epub/HC26.11.310-HBM-Bandwidth-Kim-Hynix-Hot%20Chips%20HBM%202014%20v7.pdf (visited on 07/12/2024).

[14] J. S. Meena, S. M. Sze, U. Chand, and T.-Y. Tseng, “Overview of emerging nonvolatile memory technologies”, *Nanoscale research letters*, vol. 9, pp. 1–33, 2014.

References VII

[15] M. Sochacki, “DVD-R, beschreib- und lesbare Seite. Der bereits beschriebene Innenbereich ist aufgrund seiner veränderten Reflexionseigenschaften erkennbar”, (2007), [Online]. Available: https://de.wikipedia.org/wiki/DVD#/media/Datei:DVD.png.

[16] H. Lee, M. Kim, H. Kim, H. Kim, and H.-J. Lee, “Integration and boost of a read-modify-write module in phase change memory system”, *IEEE Transactions on Computers*, vol. 68, no. 12, pp. 1772–1784, 2019. DOI: 10.1109/TC.2019.2933826.

References VIII

[17] P. Guo, A. M. Sarangan, and I. Agha, "A review of germanium-antimony-telluride phase change materials for non-volatile memories and optical modulators", *Applied Sciences*, vol. 9, no. 3, 2019, ISSN: 2076-3417. DOI: `10.3390/app9030530`. [Online]. Available: `https://www.mdpi.com/2076-3417/9/3/530`.

[18] R. Bläsing, A. A. Khan, P. C. Filippou, *et al.*, "Magnetic racetrack memory: From physics to the cusp of applications within a decade", *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1303–1321, 2020. DOI: `10.1109/JPROC.2020.2975719`.