Memory Technology Overview

SRAM (Static RAM)

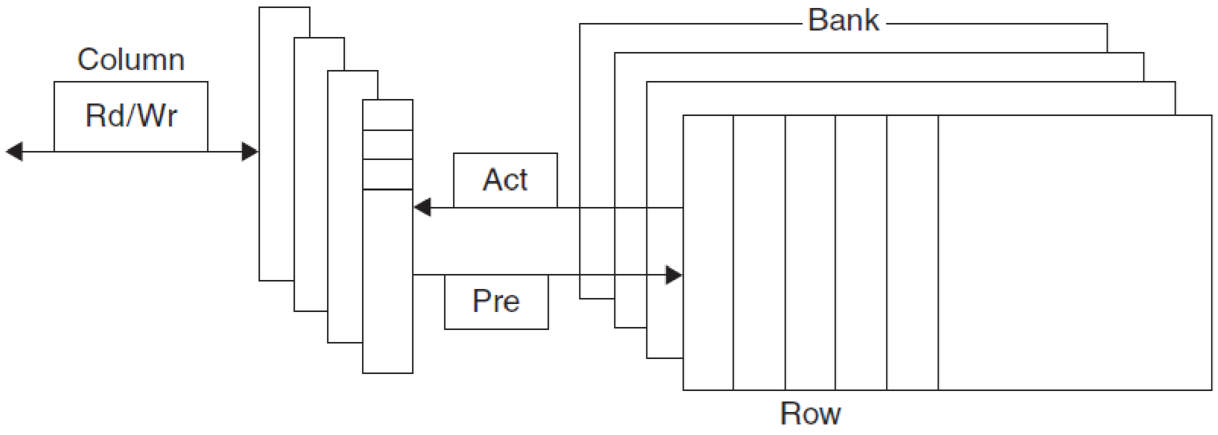
- Used primarily for cache memory
- Six transistors per bit
- No refresh needed
- Access time close to cycle time
- Minimal standby power
- Faster but more expensive than DRAM

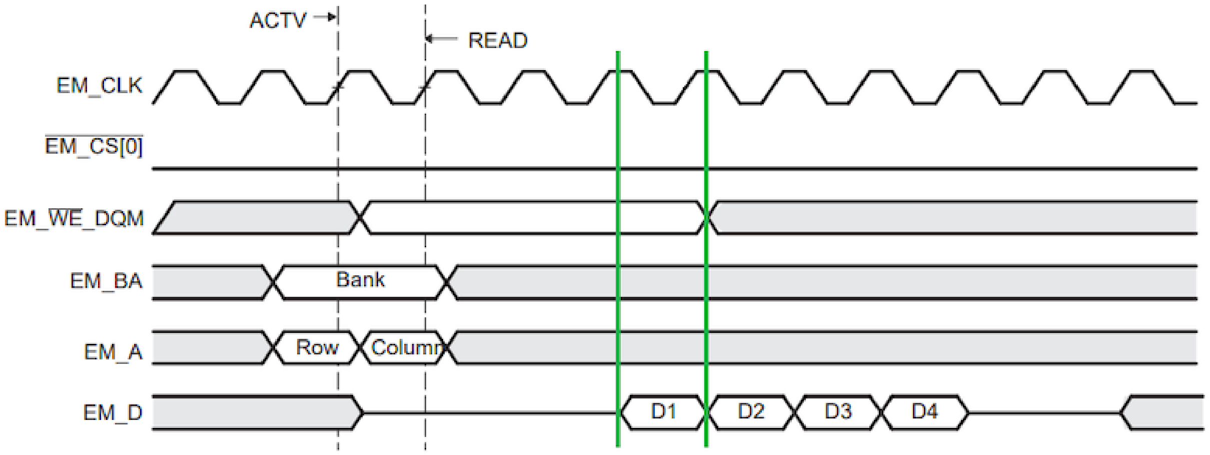
DRAM (Dynamic RAM)

- Used for main memory
- Single transistor per bit
- Requires periodic refresh
- Reading destroys data, must be restored
- Multiplexed address lines (RAS/CAS)
- Higher capacity, lower cost than SRAM

DRAM Organization

- Modern DRAMs are organized in banks, typically four for DDR3.
- Each bank consists of rows that transfer to a buffer when activated.
- Column addresses then access specific data within the row buffer.
- Commands like precharge (Pre) and activate (Act) control bank operations, with all transfers synchronized to a clock signal.
- This organization allows for higher bandwidth through multiple banks operating independently, similar to interleaving in cache design.





DRAM Performance Evolution

1

1980s: Basic DRAM

64K-4M bit capacity with 150-80ns row access times. Column access times improved from 75ns to 20ns. Cycle times ranged from 250ns to 165ns.

2

1990s: SDRAM Introduction

16M-128M bit capacity. Synchronous interface added with clock signal. Row access improved to 50-70ns, column access to 10-15ns.

3

2000s: DDR Era

256M-2G bit capacity. Double Data Rate transfers on both clock edges. Multiple banks added. Row access 30-45ns, column access dropping to 2.5-5ns.

4

2010s: DDR3/DDR4

4G-8G bit capacity. Row access times of 24-36ns, column access times below 1ns. Cycle times reduced to 31-37ns. DDR4 introduced with further improvements.

Row access time (latency) has improved at approximately 5% per year, while column access time (bandwidth) has improved at more than twice that rate.

SDRAM Innovations

Synchronous Interface

Added clock signal to eliminate overhead of synchronizing with memory controller for each transfer. Enables burst mode for higher bandwidth.

Double Data Rate (DDR)

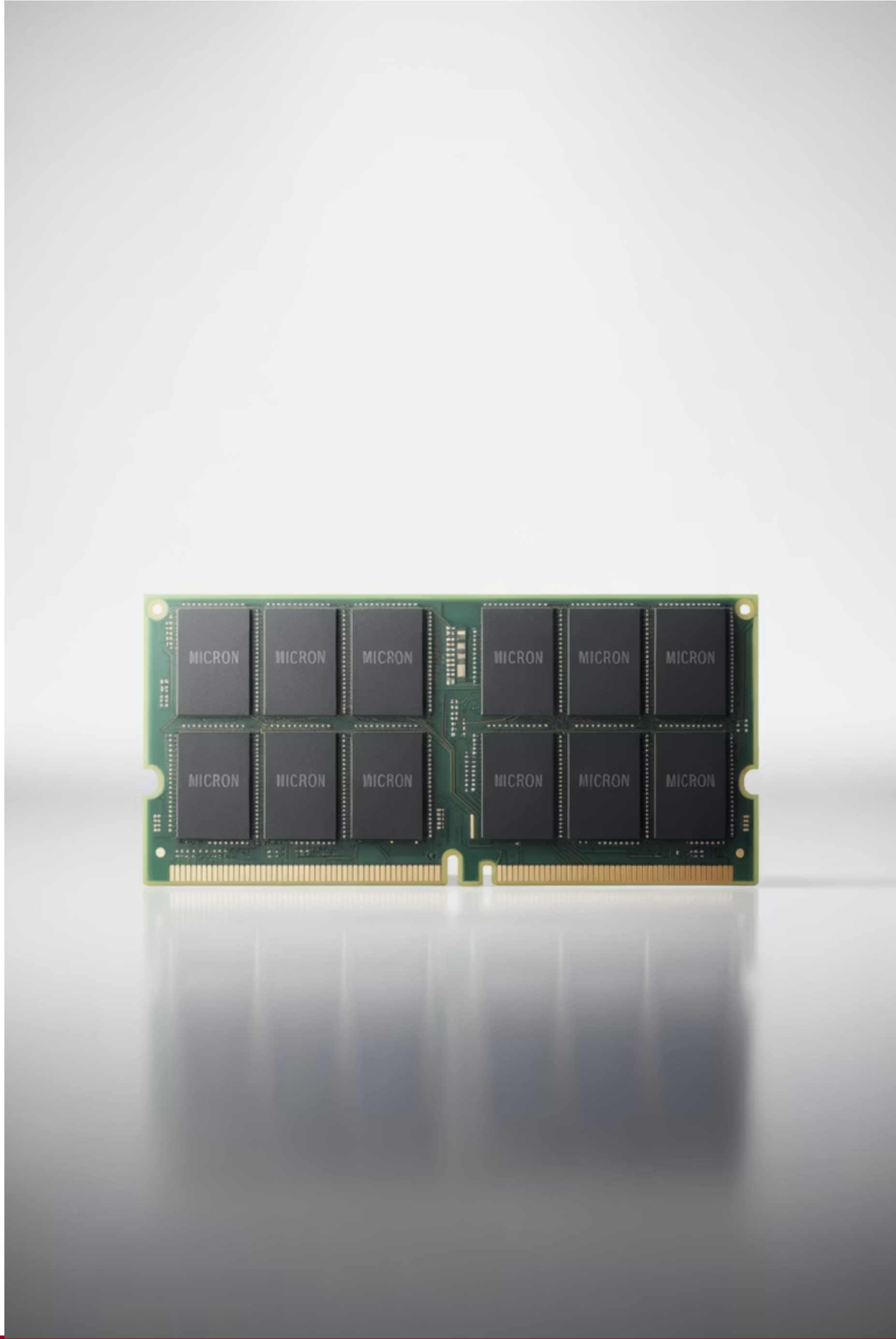
Transfers data on both rising and falling clock edges, doubling peak data rate without increasing clock frequency.

Multiple Banks

Divides DRAM into independent blocks (2-8 in DDR3) that can operate simultaneously, increasing effective bandwidth and helping with power management.

Wider Data Paths

Increased from 4-bit to 16-bit buses to deliver more data per access without increasing system size.



DDR SDRAM Generations

Standard	Clock Rate	Transfers/sec	Bandwidth/DIMM
DDR	133-200 MHz	266-400M	2128-3200 MB/s
DDR2	266-400 MHz	533-800M	4264-6400 MB/s
DDR3	533-800 MHz	1066-1600M	8528-12,800 MB/s
DDR4	1066-1600 MHz	2133-3200M	17,056-25,600 MB/s

Each generation has reduced voltage (DDR: 2.5V, DDR2: 1.8V, DDR3: 1.5V, DDR4: 1-1.2V) while increasing clock rates and bandwidth. DDR4 was scheduled for production in 2012-2014, with DDR5 following around 2014-2015.

DIMM names (like PC2100) come from peak bandwidth: 133 MHz × 2 × 8 bytes = 2100 MB/sec.

Specialized Memories

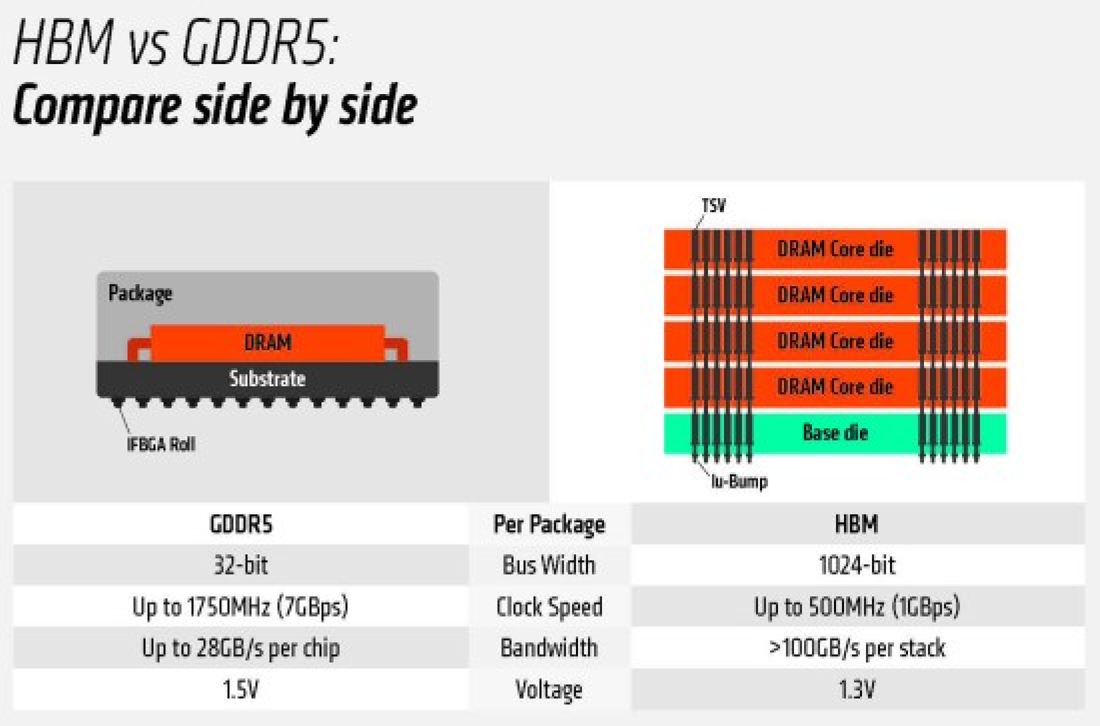
Graphics DRAM (GDDR)

- Based on SDRAM designs but optimized for GPUs
- Wider interfaces (32-bit vs 4-16 bit)
- Higher maximum clock rates
- Directly soldered to boards (not in DIMMs)
- 2-5x bandwidth per chip vs DDR3

Flash Memory

- Non-volatile (retains data without power)
- Block-based erasure before writing
- Limited write cycles (≥100,000 per block)
- Slower than DRAM but faster than disk
- Price point between DRAM and disk
- Used in mobile devices and SSDs

High-Bandwidth Memory (HBM)

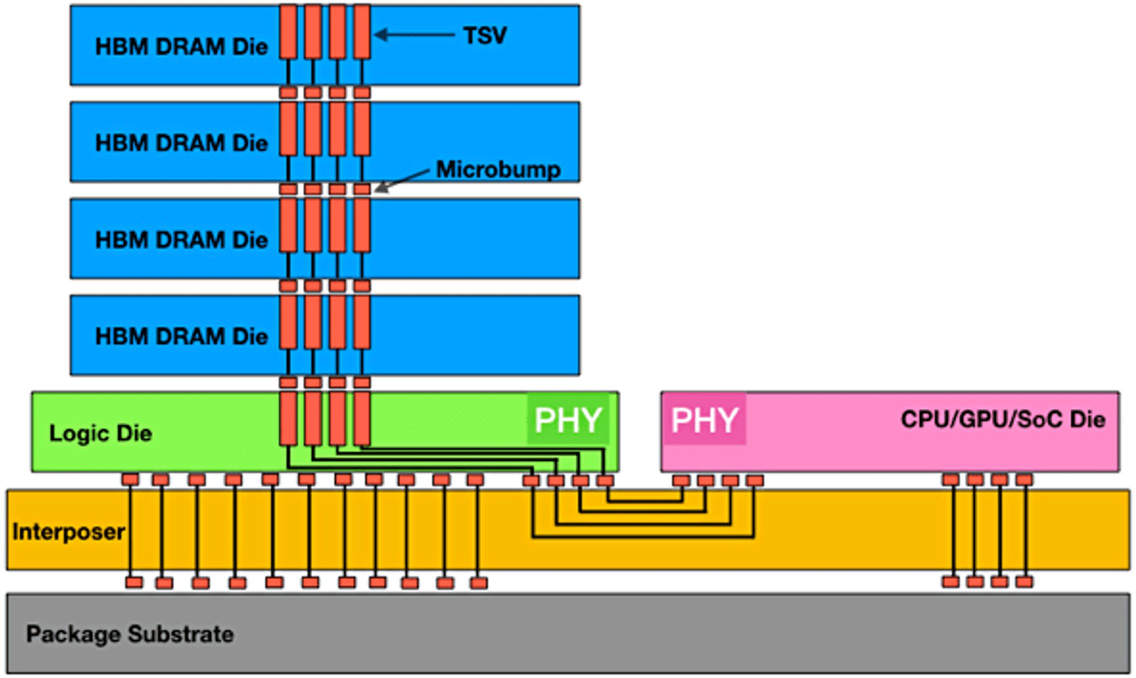


HBM vs GDDR5:

Compare side by side

Package
DRAM
Substrate
IFBGA Roll
TSV
DRAM Core die
DRAM Core die
DRAM Core die
DRAM Core die
Base die
lu-Bump

GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V



HBM DRAM Die
TSV
HBM DRAM Die
Microbump
HBM DRAM Die
HBM DRAM Die
Logic Die
PHY
PHY
CPU/GPU/SoC Die
Interposer
Package Substrate