International Journal of Advance Research in Engineering, Science & Technology(IJAREST),
ISSN(O):2393-9877, ISSN(P): 2394-2444,
Volume 2,Issue 5, May- 2015, Impact Factor:2.125

SURVEY OF MEMORY HIERARCHY AND HYBRID MEMORY CUBE

[1] Prof. Arun Tigadi, [2]Rahul C Kodaganur,[3]Prof Pramod Naik ,[4]Dr. Hansraj Guhilot

[1,2] Department of Electronics and Communication Engineering, KLE's Dr. M S Sheshgiri College of Engineering and Technology Belagavi, Karnataka
[3]Department of Electronics & Communication Engineering, VCET, Puttur, Karnataka, India
[4]Principal K.C.College of Engineering & Management Studies and Research ,Thane, Maharashtra,India
Email- [1]arun.tigadi@gmail.com, [2]rkodaganur@gmail.com, [3]pramodnaik40@gmail.com, [4] hansraj.g@gmail.com
Keywords: *memory hierarchy, hybrid memory cube, primary memory, cache memory, registers*

Abstract

All computers have a memory hierarchy; it is the classification of storage elements based on their capacity, cost and access times. Most modern CPUs are so fast that for most program workloads, the bottleneck is the locality of reference of memory accesses and the efficiency of the caching and memory transfer between different levels of the hierarchy. As a result, the CPU spends much of its time idling, waiting for memory I/O to complete. The main goal of memory hierarchy is to provide CPU with necessary data (and instructions) as quickly as possible. In this paper we review different types of memories in the memory hierarchy and evolution of DRAM. We also investigate the recent trends in primary memory such as 'Hybrid Memory Cube', which is a three-dimensional DRAM architecture that improves latency, bandwidth, power and density.

I. Introduction.

Memory hierarchy distinguishes each level in the "hierarchy" by response time. Since response time, complexity, and capacity are related, the levels may also be distinguished by the controlling technology.CPU registers hold the most frequently used data. Small, fast cache memories nearby the CPU act as staging areas for a subset of the data and instructions stored in the relatively slow main memory. The main memory stages data stored on large, slow hard disks.

Programs tend to access the storage at any particular level more frequently than they access the storage at the next lower level. So the storage at the next level can be slower, and thus larger and cheaper per bit. The overall effect is a large pool of memory that costs as much as the cheap storage near the bottom of the hierarchy, but that serves data to programs at the rate of the fast storage near the top of the hierarchy. A memory hierarchy pyramid is shown in fig1.

There are three major storage levels

1. Internal – Processor registers and cache.
2. Main – the system RAM and controller cards.
3. On-line mass storage – Secondary storage.

II. CPU Registers

An 8086 microprocessor has a total of fourteen registers. Eight of the registers are general purpose registers

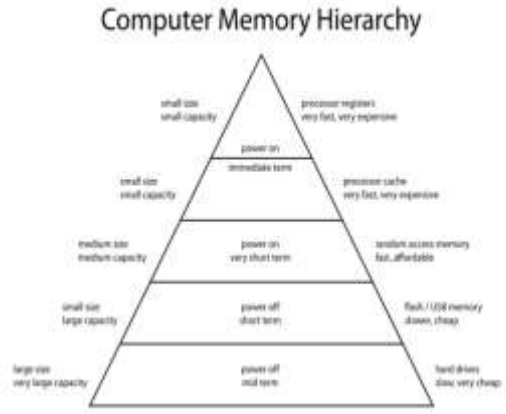

fig1. Memory Hierarchy Pyramid

the register is 16 bits long; the first four are data registers ax, bx, cx and dx.The data registers can be used as 16-bit registers or two 8-bit registers. Each 8-bit register can be used independently. The ax register may be accessed as 'ah' and 'al' (H and L refer to high-order and low-order bytes).Registers 'bx', 'cx', and 'dx' can be used in the same way.

The second four general purpose registers are index/pointer registers. They are the Stack Pointer (SP) which points to the top of the stack, Base Pointer (BP) which points the base address of the stack, Source Index (SI) this register holds the source address in an instruction and Destination Index (DI) register holds destination address in an instruction. Index registers are usually used for memory addressing

The next set of registers is four segment registers Code Segment (CS), Data Segment (DS), Stack Segment (SS) and Extra Segment (ES) register. These registers hold the base address of where a particular segment begins in memory.

The two remaining registers are the instruction pointer (IP) and the status word or flags register. Neither of these is referenced directly by program. Instruction Pointer Register is used to control which instruction the CPU executes. The IP, or program counter, is used to store the memory location of the next instruction to be executed. The CPU checks the program counter to ascertain which instruction to carry out next. It then updates the program counter to point to the next instruction.

Status (Flags) Register

It is a 16-bit register 3 bits are used as control flags and 6 bits are used as status flags, the remaining 7 are not used. The status flags are used to record specific characteristics of arithmetic and of logical instructions.

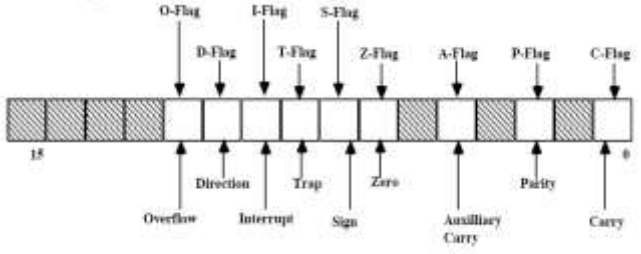

fig 2. flags register

III. Cache Memory

CPU caches are divided into 2 groups, the Level 1 and Level 2 caches, usually called L1 and L2. A L1 cache is some kind of memory which is built into the same CPU and it is the place where the CPU will first try to access. The L2 cache is another memory but instead of feeding the CPU, this one feeds the L1 cache and this way the L2 cache can be understood as a cache of the L1 cache.

L2 caches may be built the same way as the L1 caches, into the CPU but sometimes it can also be located in another chip or in a MCP (Multichip Package Module), it can also be a completely separate chip. L1 and L2 caches are built with SRAM (static RAM).

A single SRAM cell is shown below, SRAM cell consists of a bi-stable flip-flop connected to the internal circuitry by two access transistors shown below (Figure 3).

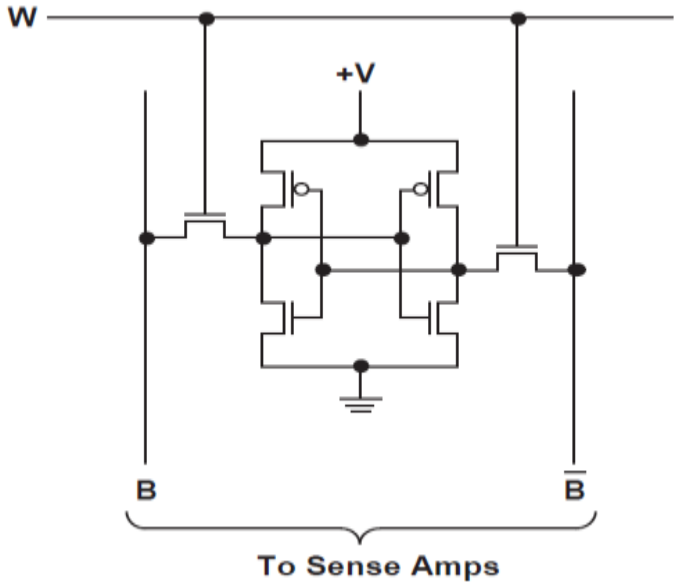

fig 3. 6T SRAM Cell

When the cell is not addressed, the two access transistors are closed and the data is kept to a stable state, latched within the flip-flop. The flip-flop needs the power supply to keep the information. The data in an SRAM cell is volatile (i.e., the data is lost when the power is removed). However, the data does not "leak away" like in a DRAM, so the SRAM does not require a refresh cycle.

The difference between L1 and L2 is the size. L1 is smaller than L2. This way the data is easier to be found in the L1 than L2, making the access much faster, if the data is not found in the L1 the data will be looked in the L2 bigger cache and if it is not there, an access to memory will be needed making the access much slower than either to L1 or L2.The way the caches are managed depends on the architecture of the processors.

IV. Main Memory

Main memory is where the programs and data are stored when the processor is actively using them. Programs and data are copied from the secondary memory into main memory where the processor can interact with them. Main memory is implemented with DRAM (Dynamic Random Access Memory) which is cheaper and slower than SRAM. A single DRAM cell along with a 64 Kbit block is shown in fig 4

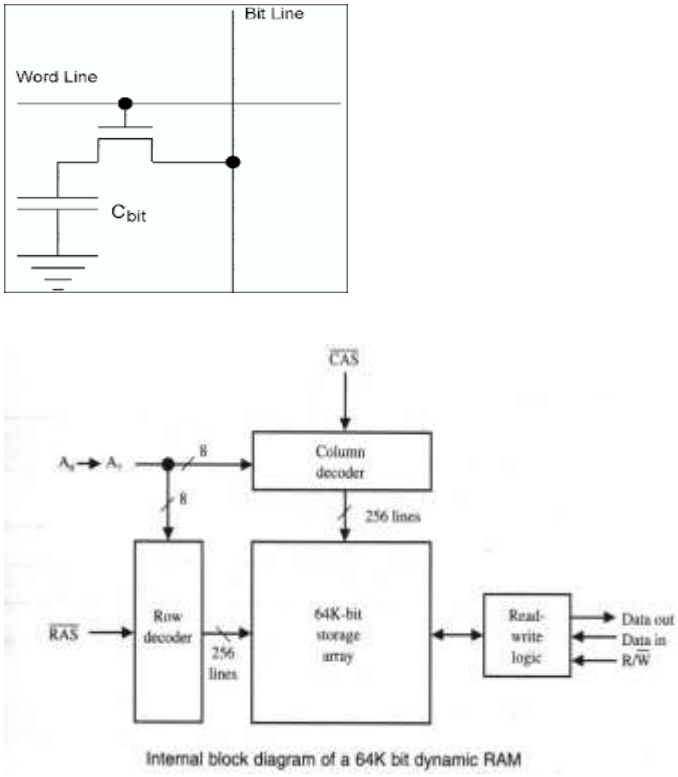

fig 4 DRAM Cell and 64K bit block

A DRAM stores each bit of data in a capacitor, the capacitor can be either charged or discharged; these two states are taken to represent logic 0 or logic 1. Since the transistor may leak a small amount of current even if it is not conducting the capacitors will slowly discharge, and the information eventually fades unless the capacitor charge is refreshed periodically. Because of this refresh requirement it is called as a dynamic memory. The DRAM write and read cycles are illustrated in fig 5 and fig 6.

Steps to read data in a DRAM cell

1. Insert the row address and make Row Address Strobe (RAS) low.
2. Make Write Enable pin high.
3. Insert the column address and make Column Address Strobe (CAS) low.
4. Read the data at I/O pin

Steps to write data in a DRAM cell

1. Insert the row address and make Row Address Strobe (RAS) low.
2. Make Write Enable pin low.
3. Insert the column address and make Column Address Strobe (CAS) low.
4. Write the data at I/O pin.

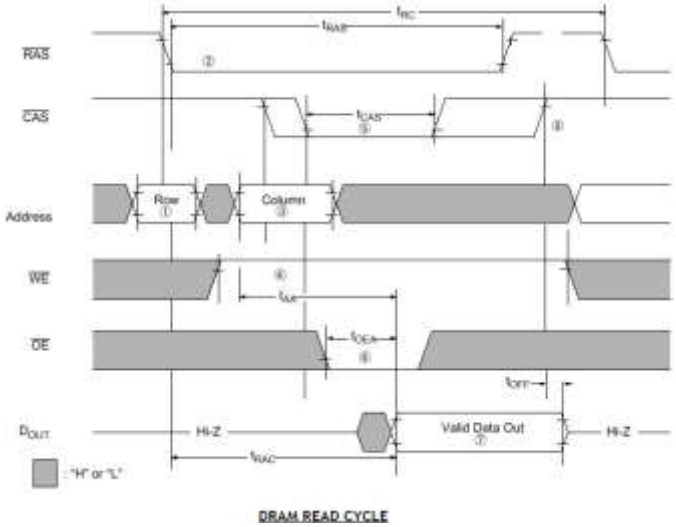

fig 5. DRAM Read Cycle

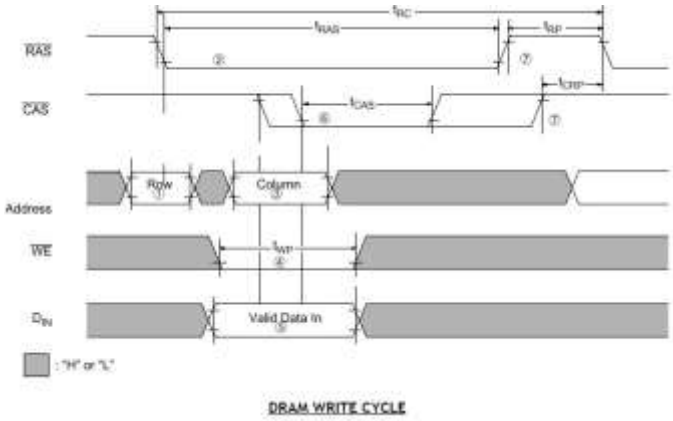

fig 6. DRAM Write Cycle

4.1 SDRAM(Synchronous Dynamic Random Access Memory)

SDRAM is a Dynamic Random Access Memory (DRAM), it is designed to synchronize itself with the timing of the CPU. This enables the memory controller to know the exact clock cycle when the requested data will be ready, so the CPU no longer has to wait between memory accesses. SDRAM is also called as SDR SDRAM (Single Data Rate SDRAM), where the I/O, internal clock and bus clock are the

same. Single Data Rate means that SDR SDRAM can only read/write one time in a clock cycle. SDRAM have to wait for the completion of the previous command to be able to do another read/write operation.

Instructions and data in personal computers tend to be read in sequential order most of the time. With a L2 Cache present, memory transactions happen as bursts of fixed sized memory blocks with continuous addresses, which is called as page mode in DRAM. The page mode access in SDRAM is called burst mode. The fig 7 shows SDRAM burst read timing.

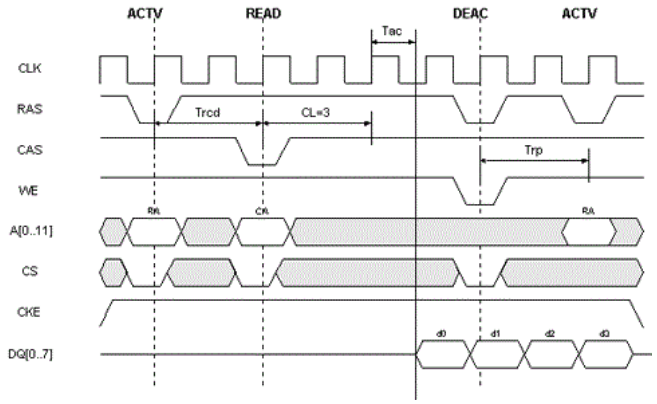

fig 7. SDRAM Burst Read Timing Diagram

First the CPU activates the row via the RAS line. After a period of time (t_{RCD}), the CAS line is activated. When the amount of time required for column access (t_{CAC}) has passed, the data appears on the output line and can be transferred on the next clock cycle. The time that has passed is approximately 50 ns for the first piece of data to become available. Subsequent transfers are then performed via burst mode (every clock cycle), or by cycling CAS if necessary, which requires an amount of time dictated by t_{CAC}, also called the CAS Latency period. For burst mode operation, the access time (t_{AC}) must be 6ns. This is so the signal can stabilize and an output operation can begin by 8 ns after the last one. The transfer of the data takes 2 ns or less, which means that the data is available every 10 ns on a burst transfer, or just in time for the next 100 MHz clock signal. Power consumption of SDRAM is 762 pj/bit.

4.2 DDR SDRAM(Double Data Rate SDRAM)

The next generation of SDRAM is DDR, which achieves greater bandwidth than the preceding single data rate SDRAM by transferring data on the rising and falling edges of the clock signal (double pumped). Effectively, it doubles the transfer rate without increasing the frequency of the clock. The transfer rate of DDR SDRAM is the double of SDR SDRAM without changing the internal clock. DDR SDRAM, is the first generation of DDR memory, the pre-fetch buffer is 2n, which is the double of SDR SDRAM. In a pre-fetch buffer architecture, when a memory access occurs to a row, the buffer grabs a set of adjacent data-words on the row and reads them out ("bursts" them) in continuous sequence on the IO pins, without the need for individual column address requests. Here n is the length of data-word.

The timing waveforms for burst read operation of DDR SDRAM are shown in fig 8.

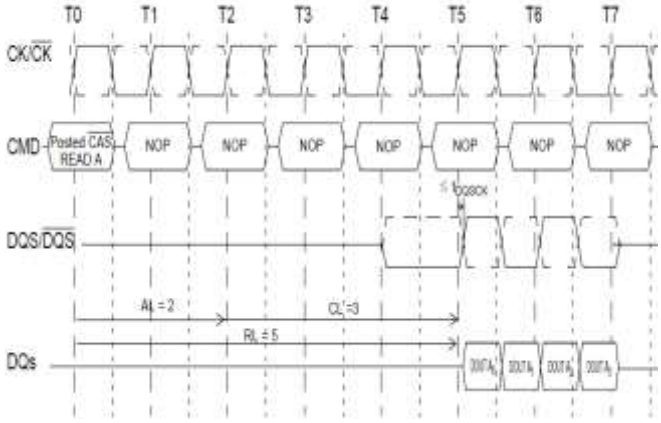

fig 8 Burst Read Operation of DDR SDRAM

The Burst Read command is initiated by having CS and CAS signal LOW while holding RAS and WE signal HIGH at the rising edge of the clock. The address inputs determine the starting column address for the burst. The delay from the start of the command to when the data from the first cell appears on the outputs is equal to the value of the read latency (RL). The data strobe output (DQS) is driven LOW 1 clock cycle before valid data (DQ) is driven onto the data bus. The first bit of the burst is synchronized with the rising edge of the data strobe (DQS). Each subsequent data-out appears on the DQ pin in phase with the DQS signal in a source synchronous manner. The RL is equal to an additive latency (AL) plus CAS latency (CL)[5].The transfer rate of DDR is between 266~400 MT/s. The operating voltage of DDR SDRAM is 2.5V. Power consumption of DDR2 SDRAM is 245 pj/bit.

4.3 DDR2 SDRAM (Double Data Rate Two SDRAM)

Its primary benefit is the ability to operate the external

data bus as in DDR SDRAM (transferring data on the rising and falling edges of the bus clock signal), DDR2 allows higher bus speed and requires lower power by running the internal clock at half the speed of the data bus. The two factors combine to produce a total of four data transfers per internal clock cycle. With data being transferred 64 bits at a time, DDR2 SDRAM gives a transfer rate of (memory clock rate) × 2 (for bus clock multiplier) × 2 (for dual rate) × 64 (number of bits transferred) / 8 (number of bits/byte). Thus with a memory clock frequency of 100 MHz, DDR2 SDRAM gives a maximum transfer rate of 3200 MB/s. Power savings are achieved primarily due to an improved manufacturing process through die shrinkage, resulting in a drop in operating voltage (1.8 V compared to DDR's 2.5 V). Power consumption is 139 pj/bit.

4.4 DDR3 SDRAM (Double Data Rate Three SDRAM)

The primary benefit of DDR3 SDRAM over DDR2 SDRAM, is its ability to transfer data at twice the rate, enabling higher bandwidth or peak data rates. With two transfers per cycle of a quadrupled clock signal, a 64-bit wide DDR3 module may achieve a transfer rate of up to 64 times the memory clock speed megahertz (MHz) in megabytes per second (MB/s). With data being transferred 64 bits at a time per memory module, DDR3 SDRAM gives a transfer rate of (memory clock rate) × 4 (for bus clock multiplier) × 2 (for data rate) × 64 (number of bits transferred) / 8 (number of bits/byte). Thus with a memory clock frequency of 100 MHz, DDR3 SDRAM gives a maximum transfer rate of 6400 MB/s.

DDR3 memory reduces 40% power consumption compared to current DDR2 modules, allowing for lower operating currents and voltages (1.5 V, compared to DDR2's 1.8 V or DDR's 2.5 V).DDR3 also adds two functions, such as ASR (Automatic Self-Refresh) and SRT (Self-Refresh Temperature). They can make the memory control the refresh rate according to the temperature variation. Power consumption DDR3 SDRAM is 52 pj/bit.

4.5 DDR4 SDRAM(Double Data Rate fourth SDRAM)

The primary advantages of DDR4 as compared to DDR3 is it includes higher module density and lower voltage requirements, coupled with higher data rate transfer speeds.DDR4 operates at a voltage of 1.2 V with a frequency between 1600 and 3200 MHz, compared to frequencies between 800 and 2400 MHz and voltage requirements of 1.5 or 1.65 V of DDR3.

DDR4 incorporates ECC memory. Error-correcting code memory (ECC memory) can detect and correct the most common kinds of internal data corruption. Parity bits are used to detect and correct the errors.DDR4 also has bus inversion technique as a power saving scheme. In this technique, the current and next state of the data bus is checked, if in the next state more than n/2 bits of a n-bit wide data bus are toggling, then the data on the bus in the next state is completed to avoid bus switching loss. Power consumption of DDR4 RAM is 39 pj/bit.

V. Hybrid Memory Cube(HMC)

Multi-core processor performance is limited by memory system bandwidth. The Hybrid Memory Cube is a three-dimensional DRAM architecture that improves latency, bandwidth, power and density. Through-silicon vias (TSVs) enable a new approach to memory system architecture. Silicon die are stacked one above the other with significantly more connections, thereby reducing the signal travel distance. As shown in Fig. below, the HMC device uses through-silicon via (TSV) technology and fine pitch copper pillar interconnect. Common DRAM logic is present separately on a high-performance logic die.[2]

The logic die is responsible for DRAM sequencing, refresh, data routing, error correction and high-speed interconnect to the host.HMC uses a simple abstracted protocol versus a traditional DRAM. The host sends read and write commands versus the traditional RAS and CAS.HMC operate at a low voltage of 1.2V, they have a large bandwidth of 128GB/s and the power consumption is 10.82pj/bit[3], which is very low as compared to DDR4 SDRAM (38.67pj/bit). The energy × bandwidth improvement is very high.

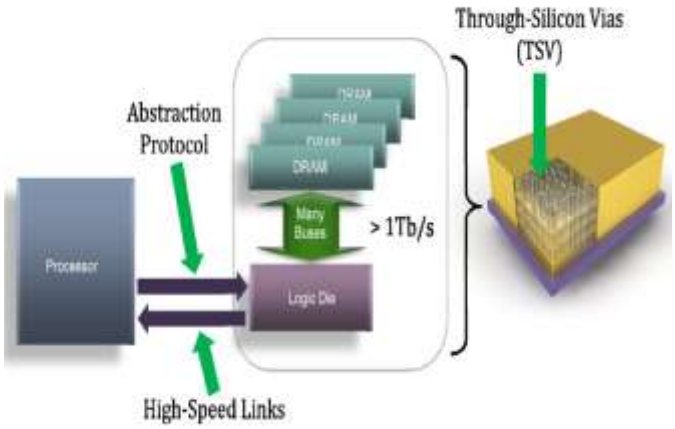

fig 9. HMC System Diagram

Memory vaults are vertical stacks of DRAM partitions (Fig. 10). The tiled architecture and large number of banks results in lower system latency under heavy load. The logic layer contains a DRAM sequencer per vault. Local, distributed control minimizes complexity. ECC is included with the data to support reliability, availability and serviceability (RAS) features. Each partition consists of 32 data TSV connections and additional command/address/ECC connections.

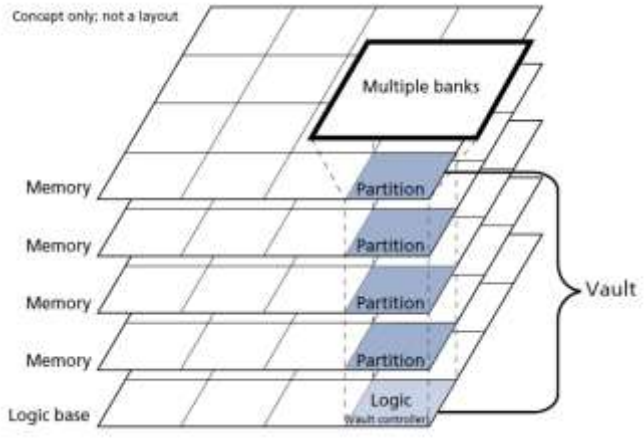

fig 9. HMC System Diagram[7]

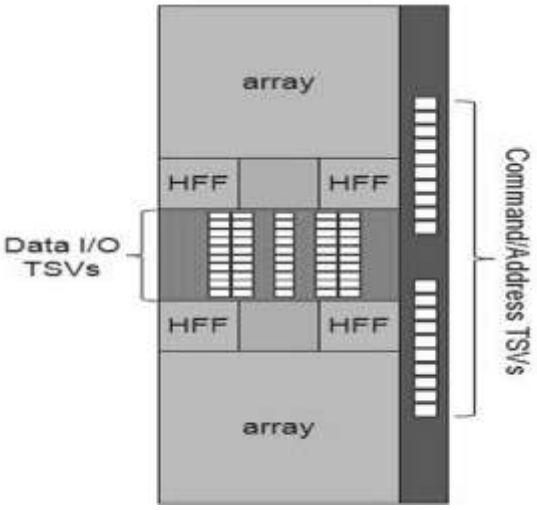

fig 11. HMC DRAM Floor Plan

References

[1] Zhiyu Liu and Volkan Kursun, et al,"High Read Stability and Low Leakage Cache Memory Cell" 1-4244-0921-7/07 2007 IEEE.
[2] Joe Jeddeloh, Brent Keeth, et al, "Hybrid Memory Cube New DRAM Architecture Increases Density and Performance" 978-1-4673-0847-2/12 IEEE 2012 Symposium on VLSI Technology Digest of Technical Papers.
[3] "Hybrid Memory Cube" by J. Thomas Pawlowski (Micron Technology)
[4] "Memory Hierarchy". Unitity Semiconductor Corporation. 16 September 2009.
[5] http://www.jedec.org/download/search/JESD79F.pdf
[6] "How Intel Plans to Transition Between DDR3 and DDR4 for the Mainstream". TechPowerUp. April 2015.
[7] HMC_Specification 1_0.pdf, Hybrid Memory Cube Consortium.