Computer Architecture

计算机体系结构

第14讲：Memory（4）

张献伟

xianweiz.github.io

DCS3013, 11/21/2022

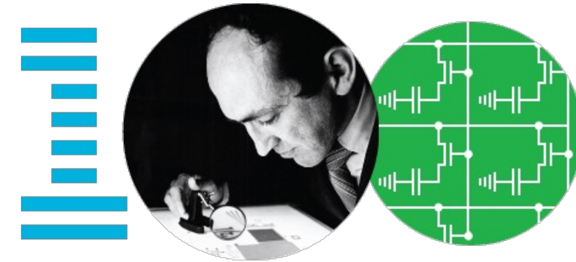
Review Questions

- Types of cache misses?
  Compulsory, capacity, conflict.
- Victim cache?
  An extra small region to hold the evicted lines from cache. Conflict.
- Miss rate vs. miss ratio?
  Misses per kilo instructions; fraction of misses out of all accesses.
- Reduce miss ratio?
  Larger capacity, higher associativity, more levels, larger blocks.
- TLB?
  Translation lookaside buffer, to hold frequently accessed pages.
- SRAM vs. DRAM?
  Static vs. dynamic, complex vs. simple, cache vs. memory.

DRAM

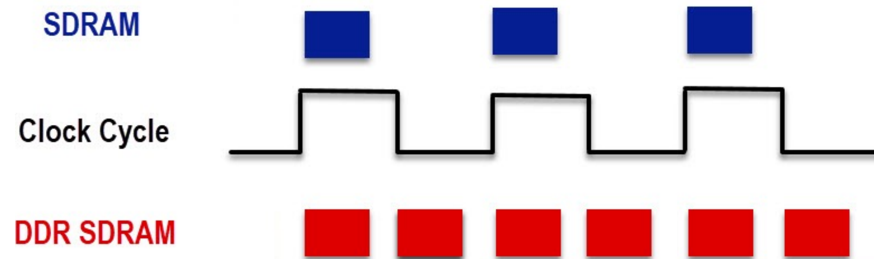
- History
  - 1966: Invented by Robert Dennard of IBM
  - 1967: DRAM patent was filed (issued 1968)
  - 1970: Intel built 1Kb DRAM chip (3T cell)
  - ~1975: 4Kb DRAM chip (1T cell)

  Dennard Scaling Law: as transistors shrank, so did necessary voltage and current; power is proportional to the area of the transistor



"I knew it was going to be a big thing, but I didn't know it would grow to have the wide impact it has today."

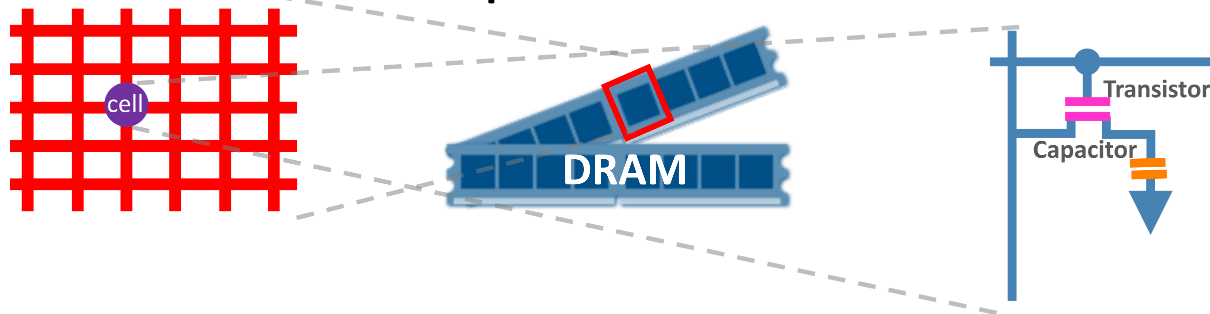
- SDRAM = DRAM with a clocked interface
- DDR SDRAM = double data rate, transfer data at both clock edges
  - DDR1 (2.5 V, 200-400 MHz)
  - DDR2 (1.8 V, 400-1066 MHz)
  - DDR3 (1.5 V, 800-2133MHz)
  - DDR4 (1.2 V, 1600-5333 MHz)
  - DDR5 (1.1 V, 3200-6400 MHz)



https://www.ibm.com/ibm/history/ibm100/us/en/icons/dram/

DRAM Structure[结构]

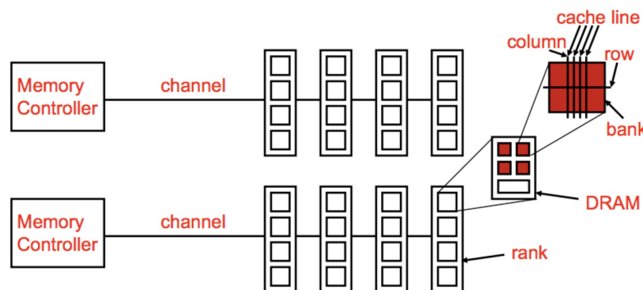
- DRAM is provided as DIMMs, which contain a bunch of chips on each side
- DRAM chip can be thought of as 2D array
- Each intersection in the array is one cell
- The cell itself is composed of 1T and 1C

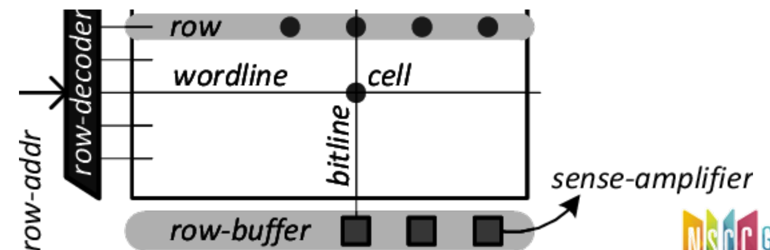


2D Array

DIMM/Chip

DRAM Cell

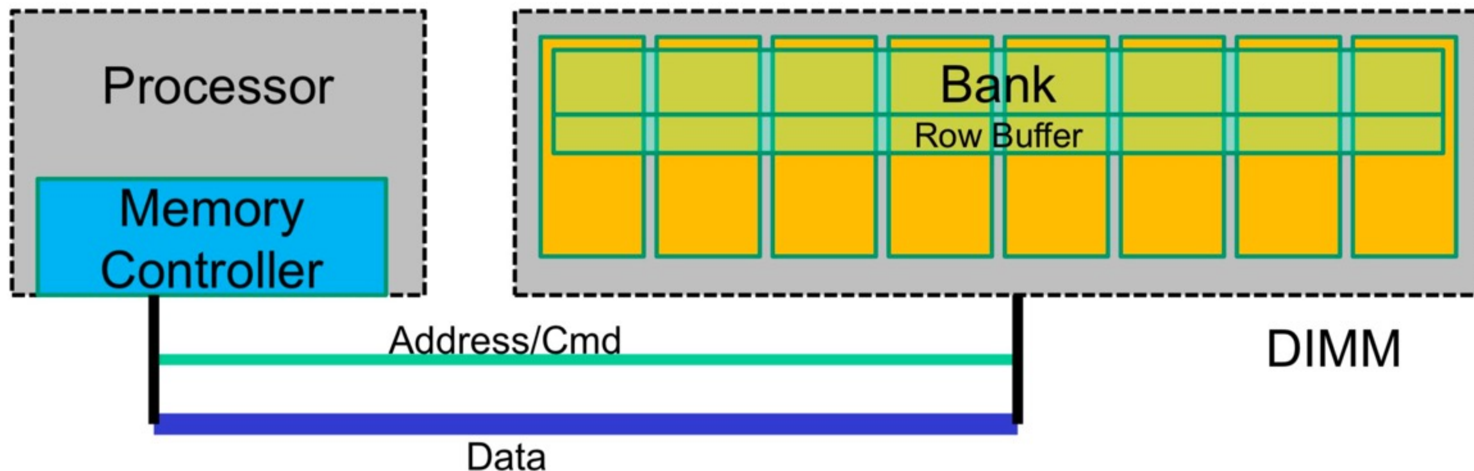




中山大學 SUN YAT-SEN UNIVERSITY

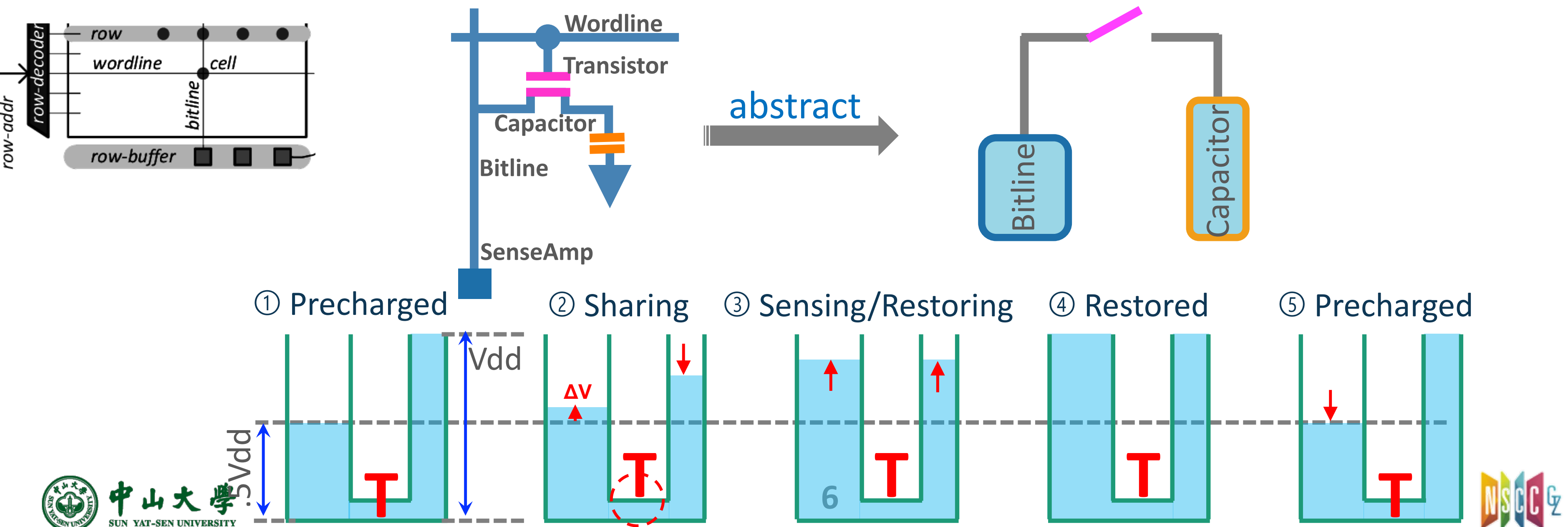
DRAM Structure (cont.)

- A rank consists of multiple (parallel) chips contributing to the same transaction
- A memory chip is organized internally as a number of banks (1-8 usually)
  - Physical bank: chip level, a portion of memory arrays
  - Logical bank: rank level, one physical bank from each chip
- Each memory bank has a "row buffer", which is non-volatile (SRAM registers)



DRAM Operations[操作]

- To read a byte (a similar process applies for writing):
  - The MC sends the row address of the byte
  - The entire row is read into the row buffer (the row is opened)
  - The MC sends the column address of the byte
  - The memory returns the byte to the controller (from the row buffer)
  - The MC sends a Pre-charge signal (close the open row)



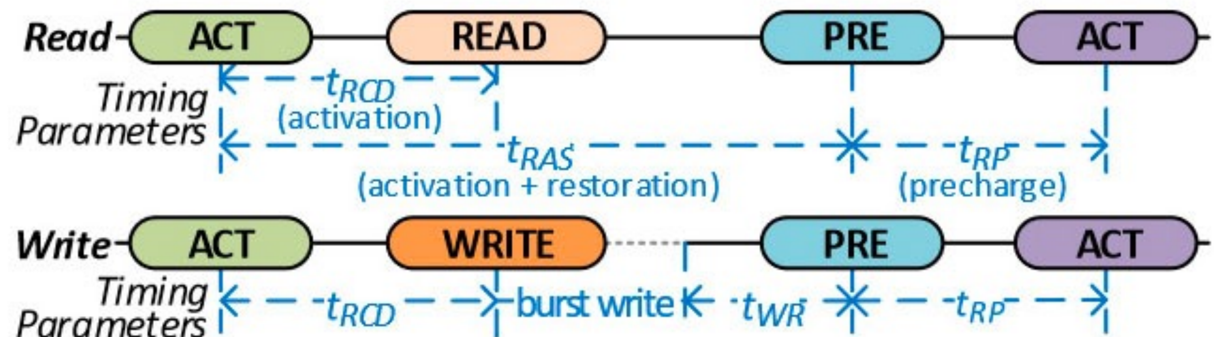
Timing Constraints[时序参数]

- Key timings
  - *tRCD*: the minimum number of clock cycles required to open a row and access a column
  - *tCAS*: number of cycles between sending a column address to the memory and the beginning of the data in response
  - *tRAS*: the minimum number of clock cycles required between a row active command and issuing the precharge command
  - *tRP*: number of clock cycles taken between the issuing of the precharge command and the active command
  - *tWR*: write recovery time

RAM TIMING

16-18-18-38

CL T_{RCD} T_{RP} T_{RAS}

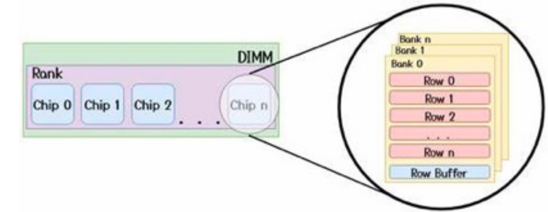


Page Mode[页模式]

- A "DRAM row" is also called a "DRAM page"
  - Usually larger than the OS page, e.g., 8KB vs. 4KB
- Row buffers act as a cache within DRAM
- Open page
  - Row buffer hit: ~20 ns access time (must only move data from row buffer to pins)
  - Row buffer conflict: ~60 ns (must first precharge the bitlines, then read new row, then move data to pins)
- Closed page
  - Empty row buffer access: ~40 ns (must first read arrays, then move data from row buffer to pins)
  - Steps
    - Activate command opens row (placed into row buffer)
    - Read/write command reads/writes column in the row buffer
    - Precharge command closes the row and prepares the bank for next access

Page Mode[页模式]

- A “DRAM row” is also called a “DRAM page”
  - Usually larger than the OS page, e.g., 8KB vs. 4KB
- Row buffers act as a cache within DRAM



- Open page
  - Row buffer hit: ~20 ns access time (must only move data from row buffer to pins)
  - Row buffer conflict: ~60 ns (must first precharge the bitlines, then read new row, then move data to pins)
- Closed page
  - Empty row buffer access: ~40 ns (must first read arrays, then move data from row buffer to pins)
  - Steps
    - Activate command opens row (placed into row buffer)
    - Read/write command reads/writes column in the row buffer
    - Precharge command closes the row and prepares the bank for next access

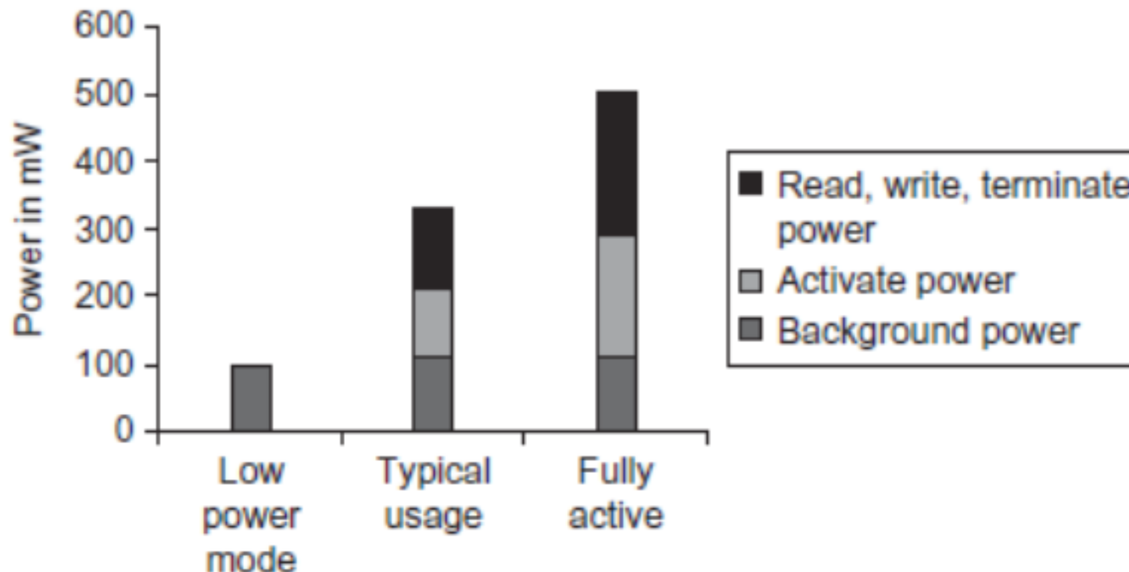
DRAM Bandwidth[带宽]

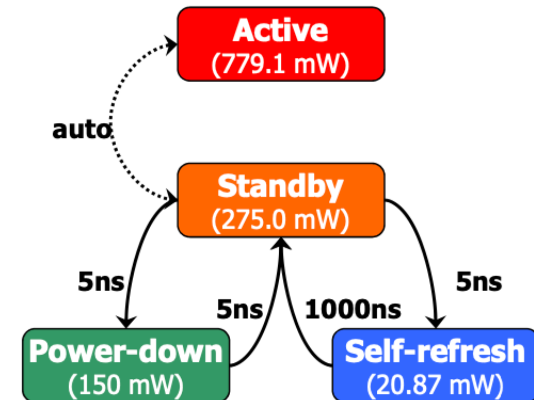
- Reading from a cell in the core array is a very slow process
  - DDR: Core speed = ½ interface speed
  - DDR2/GDDR3: Core speed = ¼ interface speed
  - DDR3/GDDR4: Core speed = ⅛ interface speed
  - … likely to be worse in the future
- Calculation: *transfer_rate * interface_width*
  - Example: 266 MT/s * 64b = 2128 MB/s

Standard	I/O clock rate	M transfers/s	DRAM name	MiB/s/DIMM	DIMM name
DDR1	133	266	DDR266	2128	PC2100
DDR1	150	300	DDR300	2400	PC2400
DDR1	200	400	DDR400	3200	PC3200
DDR2	266	533	DDR2-533	4264	PC4300
DDR2	333	667	DDR2-667	5336	PC5300
DDR2	400	800	DDR2-800	6400	PC6400
DDR3	533	1066	DDR3-1066	8528	PC8500
DDR3	666	1333	DDR3-1333	10,664	PC10700
DDR3	800	1600	DDR3-1600	12,800	PC12800
DDR4	1333	2666	DDR4-2666	21,300	PC21300

Memory Power[功耗]

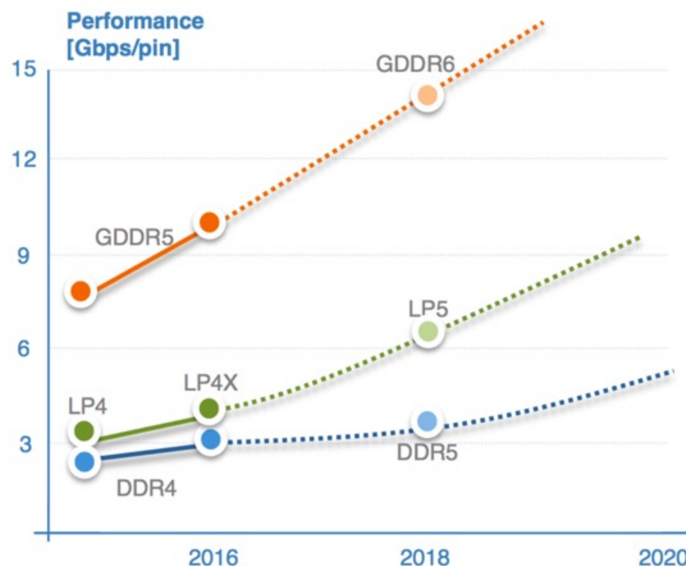
- Dynamic + static[动态和静态]
  - read/write + standby
- Reduce power[降低功耗]
  - Drop operating voltage
  - Power-down mode: disable the memory, except internal automatic refresh

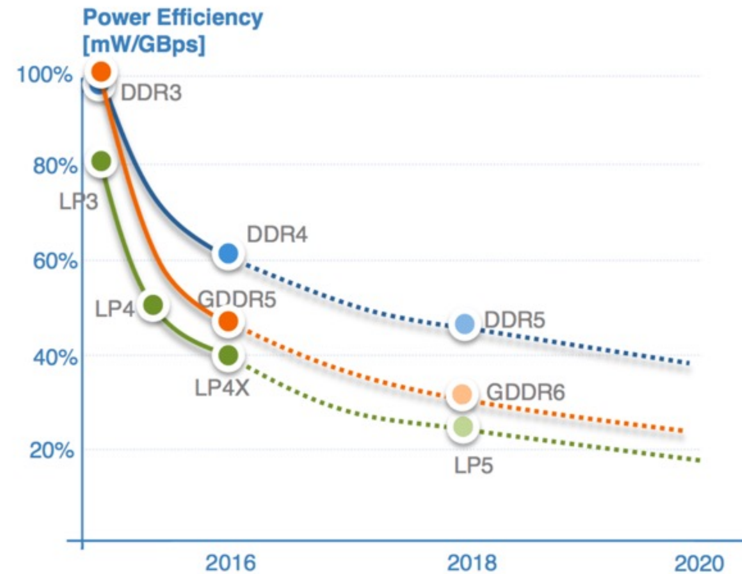




DRAM Variants[变种]

- DDR
  - DDR3: 1.5V, 800MHz, 64b → 1.6G*64b = 12.8GB/s
- GDDR: graphics memory for GPUs
  - GDDR5: based on DDR3, 8Gb/s, 32b → 8G*32b = 32GB/s
- LPDDR: low power DRAM, a.k.a., mobile memory
  - Lower voltage, narrower channel, optimized refresh



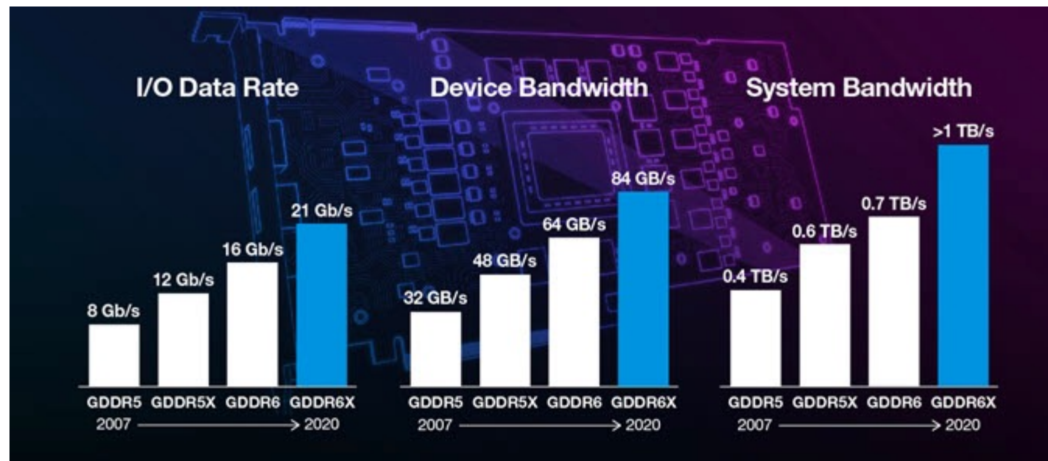


Source: ISCA2016, Samsung

SAMSUNG

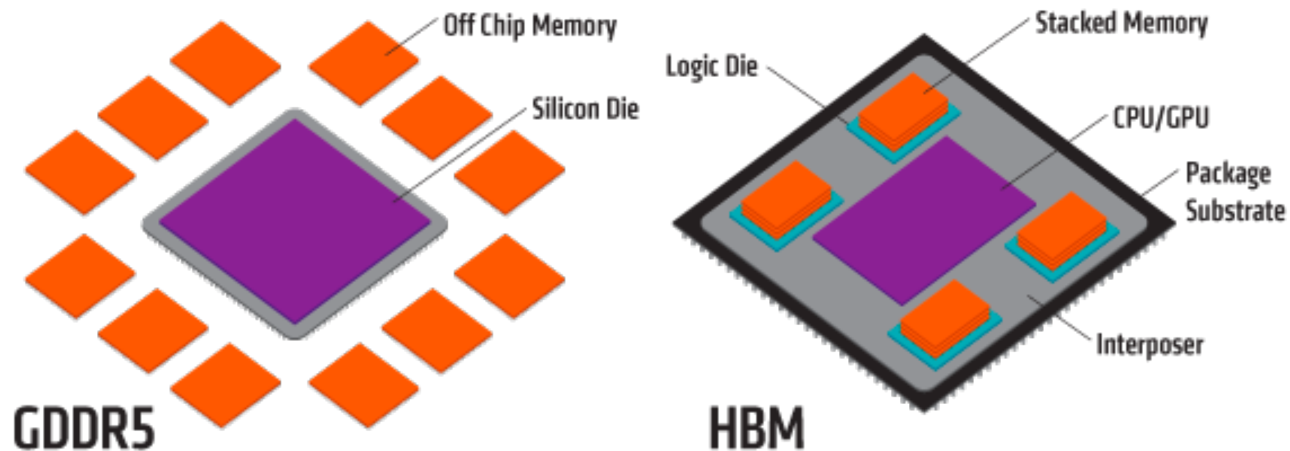
DDR5 & GDDR6

- DDR
  - DDR4: 1-1.2V, 1333MHz, 64b → 21.3GB/s x 4 = 85.2GB/s
  - DDR5: 1.1V, 6.4Gbps, 64b → 51.2GB/s x 4 = 204.8GB/s
- GDDR
  - GDDR5: 8Gb/s(~7), 256b, 224GB/s, 12GB, GTX 980
  - GDDR5X: 12Gb/s(~10), 256b, 320GB/s, 8GB, GTX 1080
  - GDDR6: 16Gb/s(~14), 256b, 448GB/s, 10GB, RTX 2080
  - GDDR6X: 21Gb/s (~19), 320b, 760GB/s, 10GB, RTX 3080
  - GDDR6X: 21Gb/s (~21), 384b, 1008GB/s, 24GB, RTX 4090



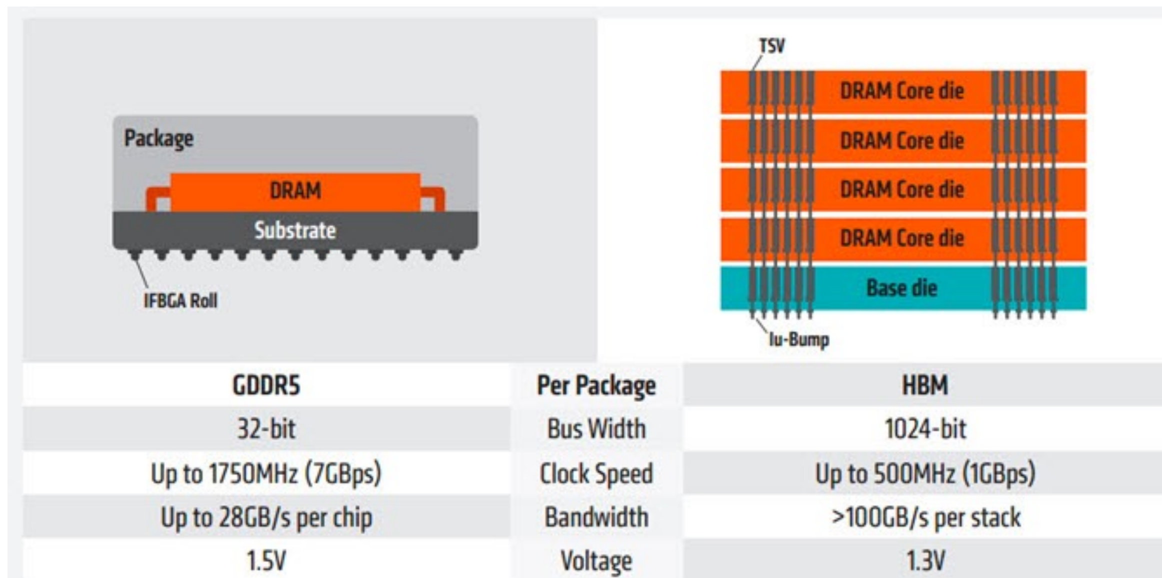
Stacked DRAMs[堆叠]

- Stacked DRAMs in same package as processor
  - High Bandwidth Memory (HBM)
- HBM consumes less power and still maintains significantly higher bandwidth in a small form factor
  - To keep the TDP target low, HBM's clock speed is limited to 1GBPs but, it makes up for it with its 4096 bits of the memory bus



HBM[高带宽内存]

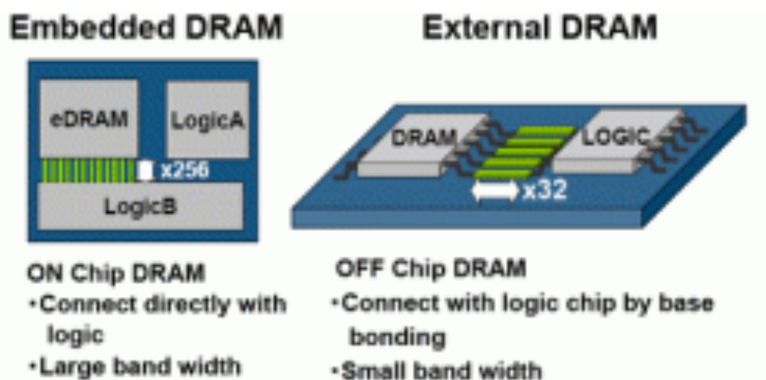
- A normal stack consist of four 4 DRAM dies on a base die and has two 128-bit channels per DRAM die
  - Making 8 channels in total which results in a 1024-bit interface
  - 4 HBM stacks gives a width of 4 * 1024 = 4096b, 1Gb/s
  - Bandwidth: 4096b * 1Gb/s = 512GB/s
- Nvidia Tesla P100: HBM2, 4096b, 16GB, 732.2GB/s
- Nvidia Tesla A100: HBM2e, 5120b, 40GB, 1555GB/s



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

eDRAM[嵌入式]

- eDRAM: embedded DRAM
  - DRAM integrated on the same die with ASIC/logic
- No pin limitations
  - Can access using a wide on-chip buses
- System power savings
  - Avoids off-chip I/O transfers

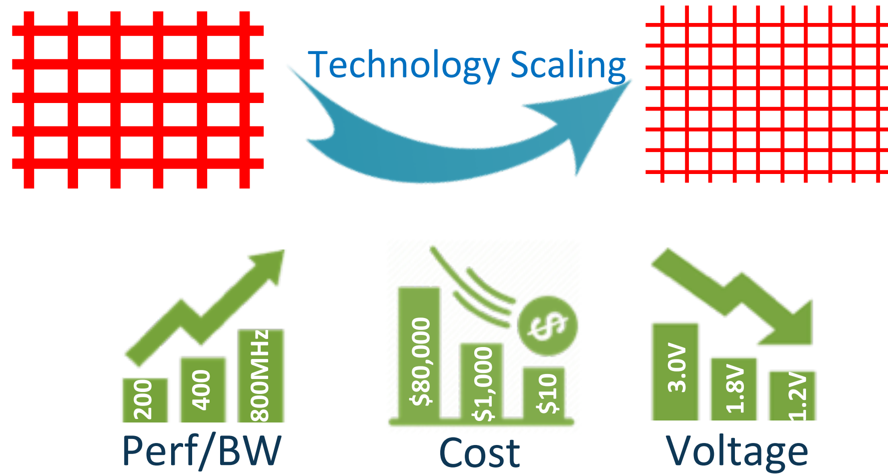


Use of eDRAM in various products

Product name	Amount of eDRAM
IBM z15	256+ MB
IBM's System Controller (SC) SCM, with L4 cache for the z15	960 MB
Intel Haswell, Iris Pro Graphics 5200 (GT3e)	128 MB
Intel Broadwell, Iris Pro Graphics 6200 (GT3e)	128 MB
Intel Skylake, Iris Graphics 540 and 550 (GT3e)	64 MB
Intel Skylake, Iris Pro Graphics 580 (GT4e)	64 or 128 MB
Intel Coffee Lake, Iris Plus Graphics 655 (GT3e)	128 MB
PlayStation 2	4 MB
PlayStation Portable	4 MB
Xbox 360	10 MB
Wii U	32 MB

https://wikimili.com/en/EDRAM

DRAM Scaling[缩放]



			Best case access time (no precharge)			Precharge needed
Production year	Chip size	DRAM type	RAS time (ns)	CAS time (ns)	Total (ns)	Total (ns)
2000	256M bit	DDR1	21	21	42	63
2002	512M bit	DDR1	15	15	30	45
2004	1G bit	DDR2	15	15	30	45
2006	2G bit	DDR2	10	10	20	30
2010	4G bit	DDR3	13	13	26	39
2016	8G bit	DDR4	13	13	26	39

Scaling Issues[问题]

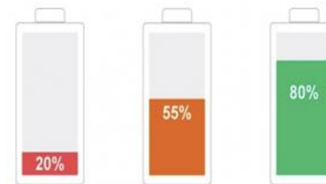
- DRAM cells are more leaky[数据流失]
  - More frequent refreshes
- Slower access[访问时延]
  - Longer sensing and restoring time
- Decreased reliability[可靠性]
  - Cross-talking noise, enlarged process variations

Less charge
higher leakage current

More Leaky

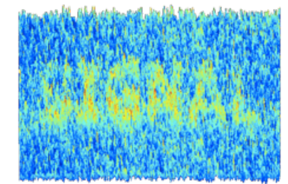
Larger resistance
Weaker signal

Longer Sensing



Larger resistance
Lower voltage

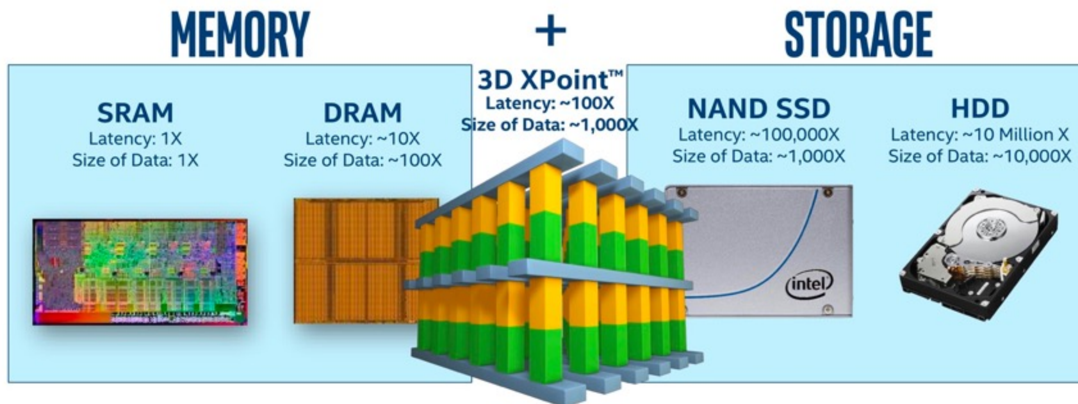
Prolonged Restore



Nearer cells
Process variations

Severer Noise

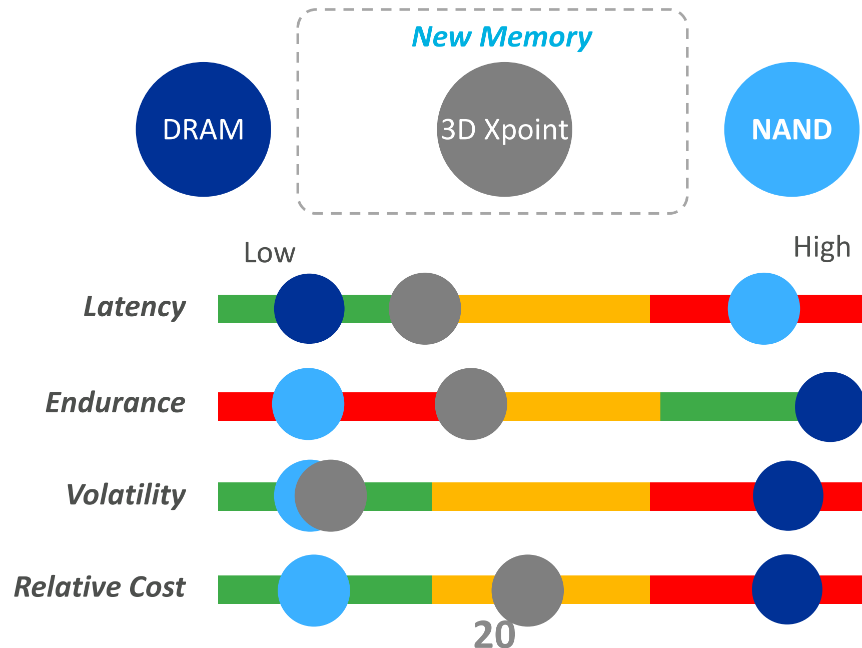
Emerging Memory[新型存储]



	Hard disk drive (HDD)	Dynamic RAM (DRAM)	NAND single-level cell (SLC) flash	Phase change RAM (PCRAM) SLC	Spin-torque transfer RAM (STT-RAM)	Resistive RAM (ReRAM)
Data retention	Y	N	Y	Y	Y	Y
Cell size (F = feature size)	N/A	6 to $10F^2$	4 to $6F^2$	4 to $12F^2$	6 to $50F^2$	4 to $10F^2$
Access granularity (Bytes)	512	64	4,192	64	64	64
Endurance (writes)	$>10^{15}$	$>10^{15}$	10^4 to 10^5	10^8 to 10^9	$>10^{15}$	10^{11}
Read latency	5 ms	50 ns	25 us	50 ns	10 ns	10 ns
Write latency	5 ms	50 ns	500 us	500 ns	50 ns	50 ns
Standby power	Disk access mechanisms	Refresh	N	N	N	N

NVM[非易失性存储]

- Numerous emerging memory candidates
  - Many fall between NAND and DRAM
- Pros and cons
  - Non-volatility with fraction of DRAM cost/bit
  - Ideal for large memory systems
  - Slower access and limited lifetime

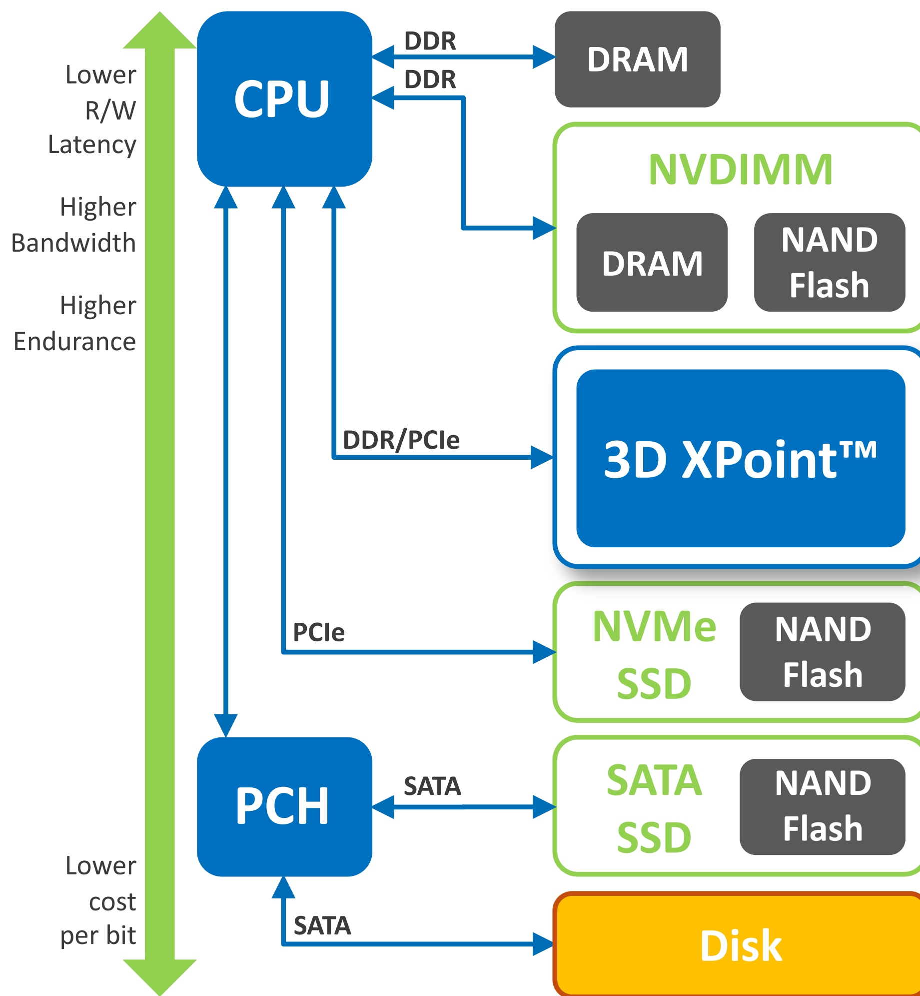


中山大學 SUN YAT-SEN UNIVERSITY

NSCCGZ

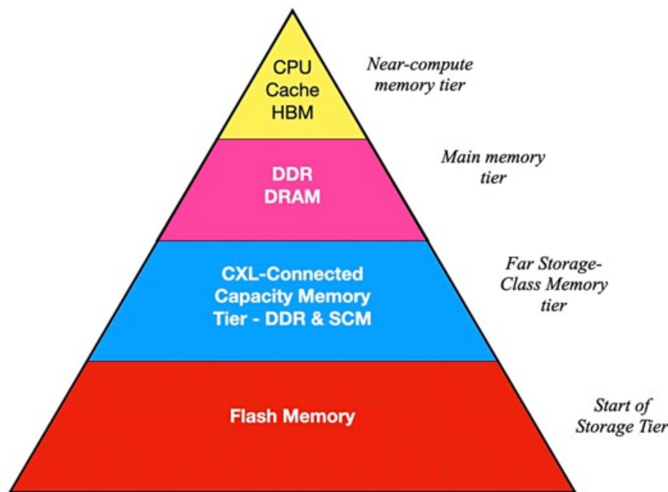
Future Memory System[未来存储系统]

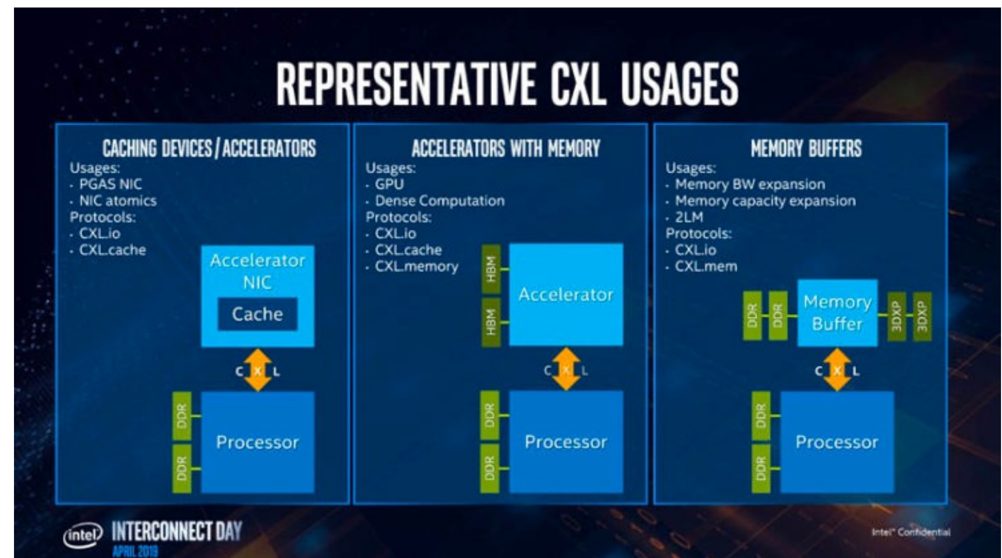
- Demands[需求]
  - Low latency
  - Large size
  - High bandwidth
  - Low power/energy
- Hybrid memory[混合]
  - DRAM + emerging
- Abstracted interface[抽象]
  - Hide device characteristics
- Changing processor-memory relationship[存算]
  - Processor-centric to memory-centric



Storage Class Memory (SCM)

- An era of very big, PB-level memory pools
- The big memory pooling is made possible by the compute express link (CXL)
- CXL is a standard for linking memory bus devices together: CPUs, GPUs, and memory (and a few other more exotic things like TPUs and DPUs).





https://www.computeexpresslink.org/