Memory

Introduction to Memory

Embedded system – Processor, Memory and I/O.

- Processing : transformation of data
- Storage : retention of data for later use
- Communication: transfer of data using buses.

Characteristics of memory: Area, Speed and Power.

Memory Classification:

- Data Storage Mode
  - Volatile: SRAM, DRAM
  - Nonvolatile:
    - Read-Write: Flash, EPROM, EEPROM
    - Read only: Mask-Programmable ROM

Types of Semiconductor memories

RAM

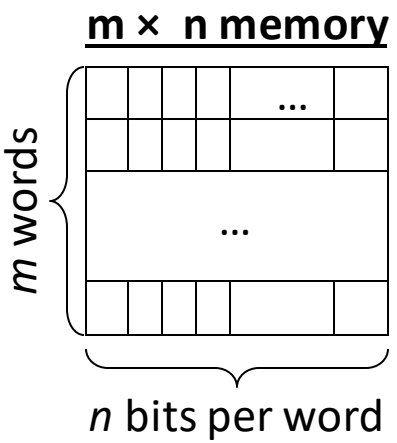
- Static RAM(SRAM)
- Dynamic RAM(DRAM)
- PSRAM (Pseudo-Static RAM)
- NV RAM(Non-Volatile RAM)

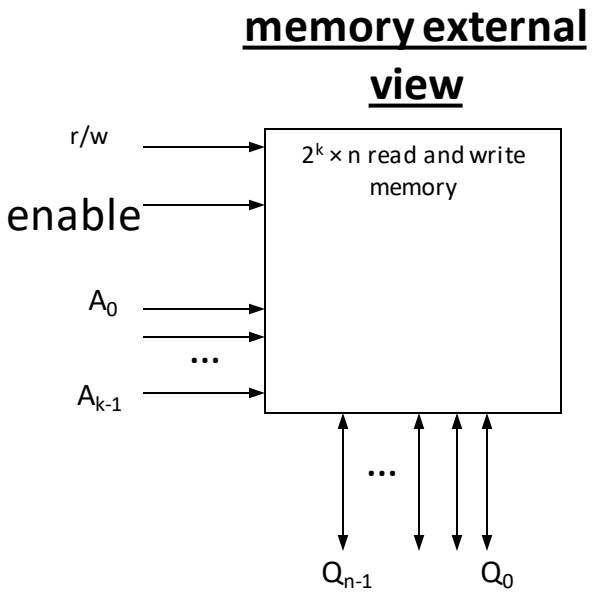
ROM

- Mask-Programmed ROM
- One-Time Programmable ROM
- Erasable Programmable ROM(EPROM)
- Electrically Erasable and Programmable ROM(EEPROM)
- Flash memory

Introduction to Memory Contd…

- Stores large number of bits
  - *m* x *n*: *m* words of *n* bits each
  - k = Log2(*m*) address input signals
  - or *m* = 2^k words
  - e.g., 4,096 x 8 memory:
    - 32,768 bits
    - 12 address input signals
    - 8 input/output data signals
- Memory access
  - r/w: selects read or write
  - enable: read or write only when asserted
  - multiport: multiple accesses to different locations simultaneously





Introduction to Memory Contd…

- Reading a memory means to retrieve the word of a particular address.
- Writing a memory means to store a word in a particular address.
- Memory access means either a read or write.
- Memory which requires both read and write needs a control input i.e., r/w.

According to Moore's Law "every 18 months memory-chip bit-capacity doubles".

ROM: a memory that a processor can only read, and which holds its stored bits even without a power source.

RAM: a memory that a processor can read and write but loses its stored bits if power is removed.

However, processor can not only read but also write to advanced ROM like EEPROM and Flash memory although the writing to this may be slow as compared to RAM.

Furthermore, advanced RAM like NVRAMs can hold their bits even when the power is removed.

Memory Write Ability and Storage Permanence

Write Ability – the manner and speed that a particular memory can be written.

Types of write ability:

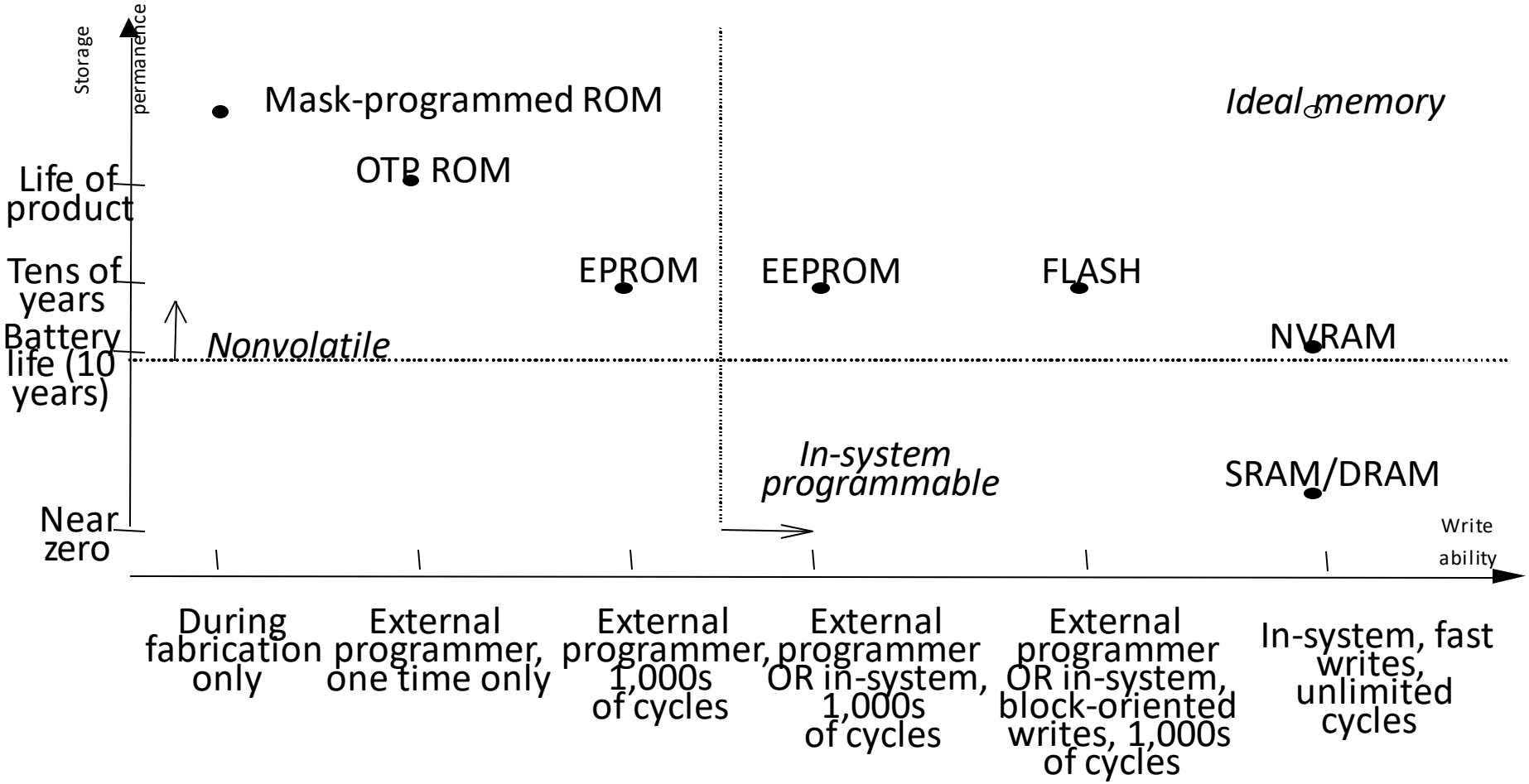
- High End Range: Processor write to memory types simply and quickly by setting such a memory's address lines, data input bit and control lines appropriately.
- Middle Range: Processor write to memory types that are slower.
- Lower End Range: Memory can be written by a special equipment called a 'programmer'. This device must apply special voltage levels to write to the memory, also known as "programming" or "burning".
- Low End Range: We have types of memory that can only have their bits stored when the memory chip itself is being fabricated.

Memory Write Ability and Storage Permanence

Storage Permanence – the ability of memory to hold its stored bits after those bits have been written.

- Low End: Memory that begins to lose its bits almost immediately after those bits are written and must be continually refreshed.
- Lower End: Memory that will hold its bits as long as power is applied to the memory.
- Middle End: Memory that can hold its bits for days, months, or even years after the memory's power source has been turned off.
- High End: Memory that will essentially never loose its bits as long as the memory chip is not damaged.

Memory Write Ability and Storage Permanence



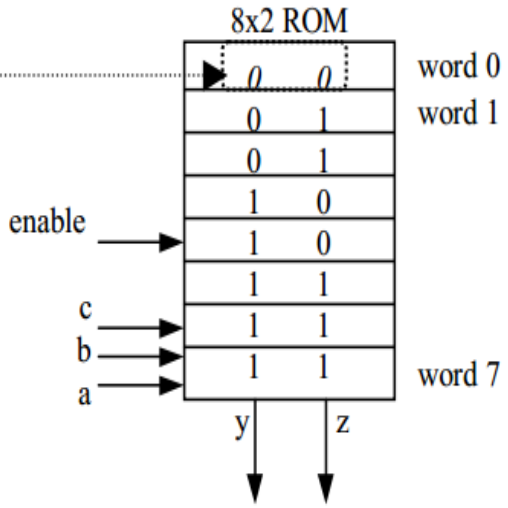
Write ability and storage permanence of memories, showing relative degrees along each axis (not to scale).

Common Memory Types

Read Only Memory – ROM

- Nonvolatile memory that can be read from but not written to by a processor.
- store a software program for a processor.
- store constant data, like large lookup tables of strings or numbers.
- to implement combinational circuit.

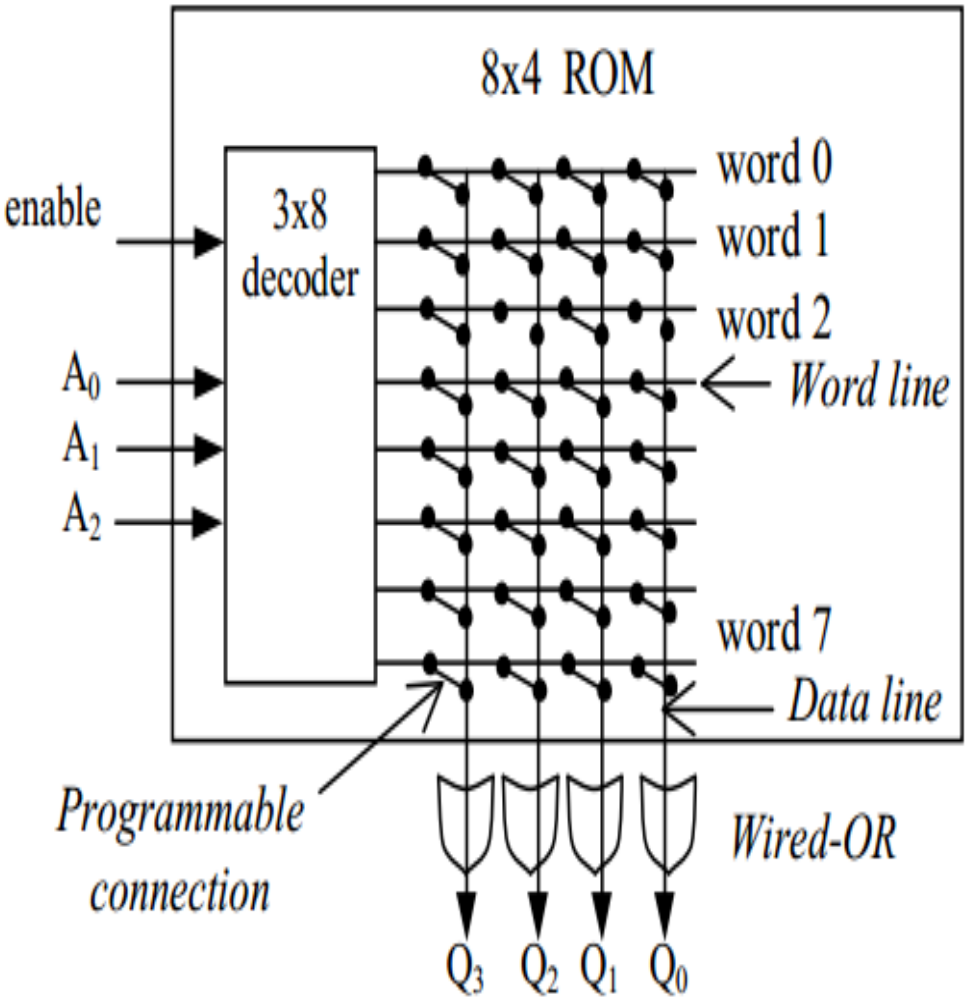
Inputs (address)			Outputs	
a	b	c	y	z
0	0	0	*0*	*0*
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	1	0
1	0	1	1	1
1	1	0	1	1
1	1	1	1	1



Common Memory Types Contd…

- Horizontal lines -> words.
- Vertical lines -> data.
- if address input is 010 then decoder sets word 2's line to 1 Data lines Q_3 and Q_1 are set to 1 because there is a "programmed" connection with word 2's line.
- Word 2 is not connected with data lines Q_2 and Q_0.
- Output is 1010.

Internal View of ROM



Common Memory Types Contd…

ROM Memory Types:

- Mask- Programmed ROM
- One-Time Programmable ROM
- Erasable Programmable ROM
- Electrically Erasable Programmable ROM
- Flash Memory

Mask-Programmed ROM

- **Connections are programmed during fabrication.**
  - Connection is programmed when the chip is being fabricated by creating appropriate set of masks.
- **Write Ability:**
  - Has extremely low write ability.
- **Storage Permanence:**
  - highest storage permanence since the stored bits will never change unless the chip is damaged.
- **Typically used for final design of high-volume systems.**

One-Time Programmable (OTP ROM or PROM)

Connections are programmed after manufactured by user.

- To program a PROM device, the user provides a file that indicates the desired ROM contents.
- A piece of equipment called a ROM programmer then configures each programmable connection according to the file.
- Uses a fuse for each programmable connection.
- ROM Programmer blows fuses by passing a large current wherever a connection should not exist. Once a fuse is blown connection can never be reestablished.

One-Time Programmable (OTP ROM or PROM)

Write Ability:

- Has lowest write ability of PROMs since they can be written only once.

Storage Permanence:

- . Has high storage permanence since their stored bits won't change unless someone reconnects the device to a programmer and blows more fuses.

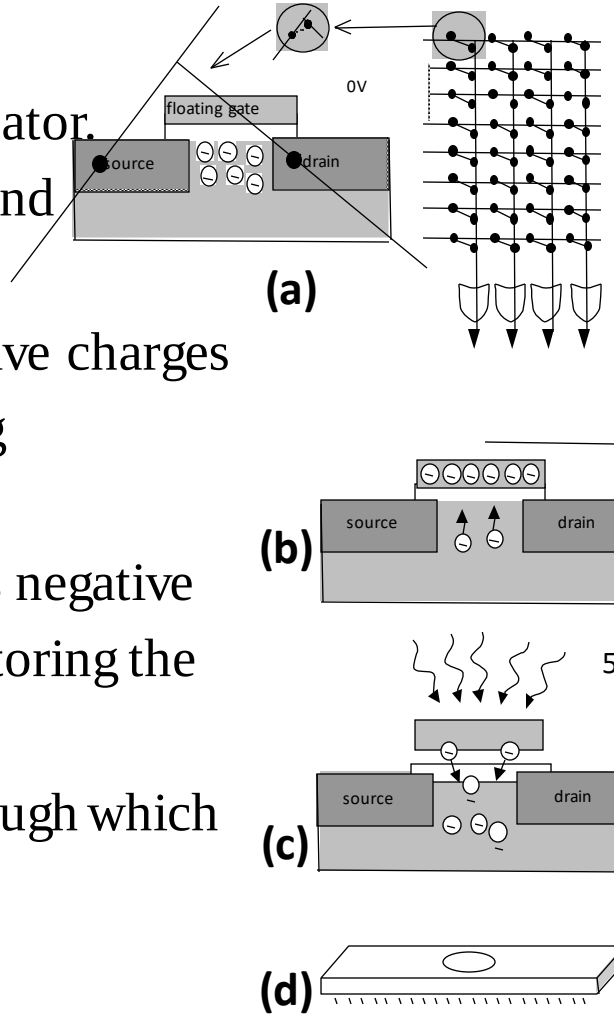
Commonly used in final products:

- This devices are better suited to prototyping and to low-volume applications than are masked-ROM
- This ROMs are also cheaper per chip than other PROMs often costing under a dollar each.

EPROM(Erasable Programmable ROM)

Programmable Component is a MOS transistor.

- Transistor has floating gate surrounded by an insulator.
- Negative charges form a channel between source and drain storing a logic 1.
- Large positive voltage(12-25) at gate causes negative charges to move out of channel and get trapped in floating gate storing a logic 0.
- Shining UV rays on surface of floating gate causes negative charges to return to channel from floating gate restoring the logic 1.
- An EPROM package showing quartz window through which UV light can pass.



EPROM(Erasable Programmable ROM)

Better Write Ability:

- Can be erased and reprogrammed thousand of times.

Reduce Storage Permanence:

- Program last about 10 years but is susceptible to radiation and electric noise.

Note: Reading an EPROM is much faster than writing, since reading does not require programming.

EEPROM(Electrically Erasable Programmable ROM) or E-squareds.

Programmed and erased electronically:

- using higher than normal voltage.
- can program and erase individual words.
- more expensive than EPROM but convenient to use.
- For example: EEPROM is used in telephones that can store commonly dialed phone numbers in memory for speed dialing.

EEPROM(Electrically Erasable Programmable ROM) or E-squareds.

Better Write Ability:

- can be in-system programmable with built-in circuit to provide higher than normal voltage.
- built-in memory controller commonly used to hide internal memory access details from memory user and provides a simple memory interface to user.
- While read access requires only tens of nanoseconds and writes may requires tens of microseconds or more.
- busy pin indicates to processor EEPROM still writing.
- can be erased and programmed tens of thousands of times.

Similar storage permanence of EPROM(10 years):

- can be erased and programmed tens of thousands of times before losing their ability to store data.

Far more convenient than EPROMs, but more expensive.

Flash Memory

Extension of EEPROM:

- same floating gate principle.
- same write ability and storage permanence.

Fast Erase

- large blocks of memory erased at once, rather than one word at a time.
- blocks typically several thousand bytes large.

Writes to single words may be slower.

- Entire block must be read, word updated, then entire block written back.

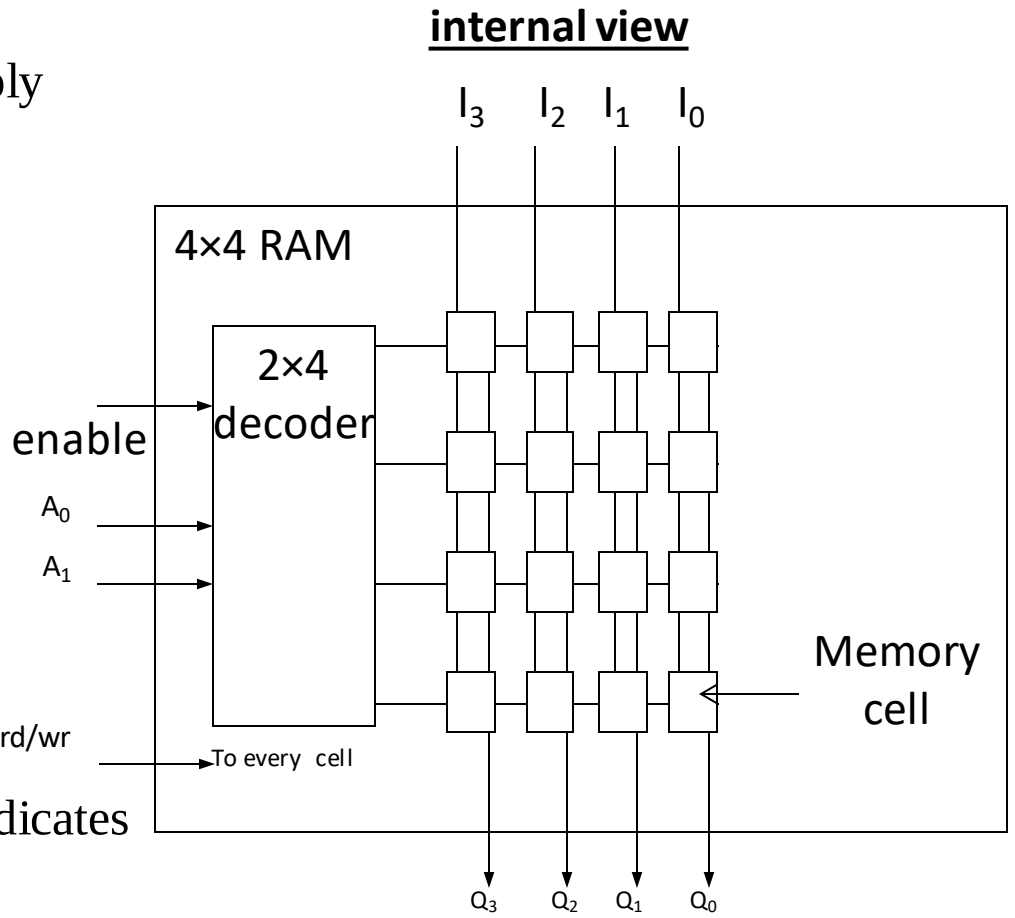
Used in embedded systems storing large data items in non volatile memory.

- e.g., digital cameras, TV set-top boxes, cell phones and medical monitoring equipment.

Note: Writing to a single word in flash may be slower than writing to single word in EEPROM.

Introduction to Read-Write Memory(RAM)

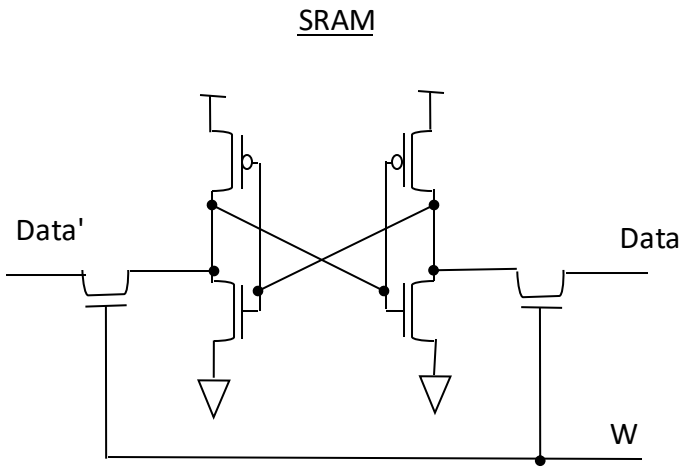
- RAM – Random Access Memory that can read and write easily.
- Writing to RAM is faster than reading.
- Typically Volatile memory
  - bit are not held without power supply
- Internal Structure of RAM
  - More complex than ROM.
  - A word consists of several memory cells each storing 1-bit.
  - Each input and output data lines connects to each cells in its column.
  - rd/wr connected to every cell.
  - When row is enabled by decoder each cell has logic that stores input data bit when rd/wr indicates write or outputs stored bit when rd/wr indicates read.



Basic Types of RAM

RAM: Static RAM

- Memory cell uses flip-flop to store bit.
- Requires 6 transistors.
- Holds data as long as power supplied.
- Typically used for high-performance parts of of a system(e.g., cache)

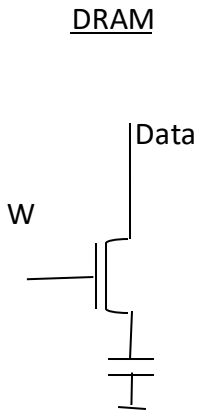


High Switching Speed.

Basic Types of RAM

RAM: Dynamic RAM

- MOS transistor and a capacitor to store a bit.
- More compact memory than SRAM.
- Refresh required due to capacitor leak
  - Word's cells refreshed when read.
- Refresh Rate 15.625.
- Slower to access than SRAM.



RAM Variations

PSRAM: Pseudo-static RAM

- DRAM with built-in memory refresh controller.
- Popular low-cost high-density alternative to SRAM.

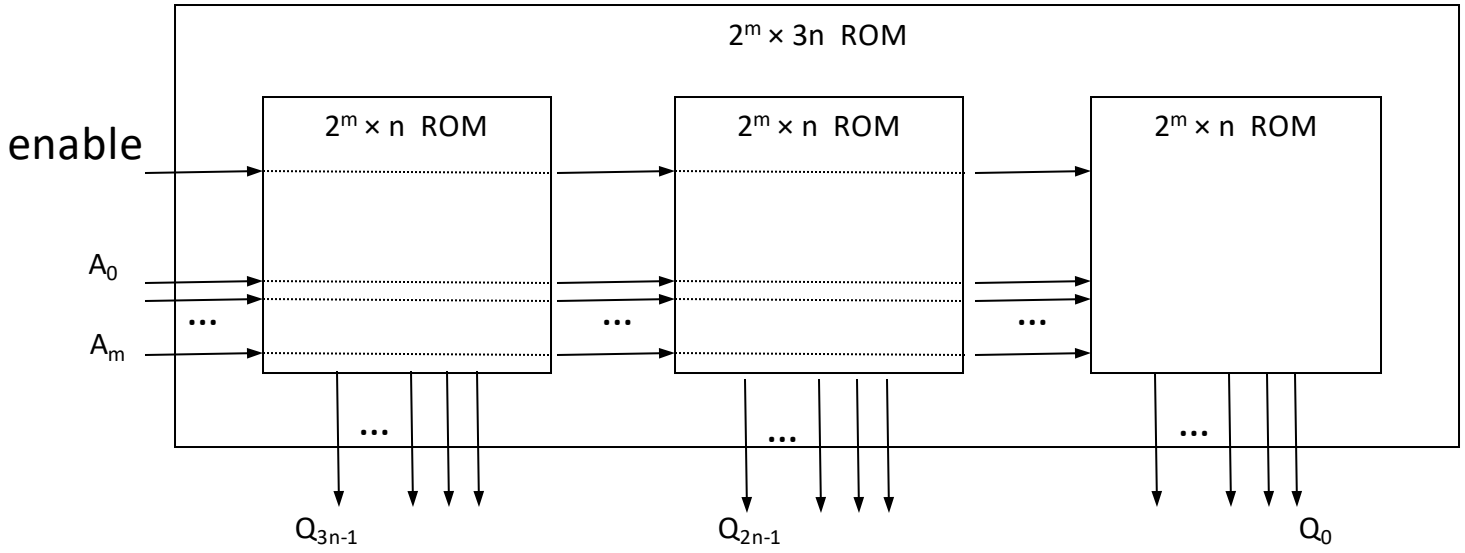
NVRAM: Nonvolatile RAM

- Holds data after external power removed
- Battery-backed RAM
  - SRAM with own permanently connected battery.
  - writes as fast as reads.
  - no limit on number of writes unlike nonvolatile ROM-based memory.
- SRAM with EEPROM or flash
  - stores complete RAM contents on EEPROM or flash before power turned off

Composing memory

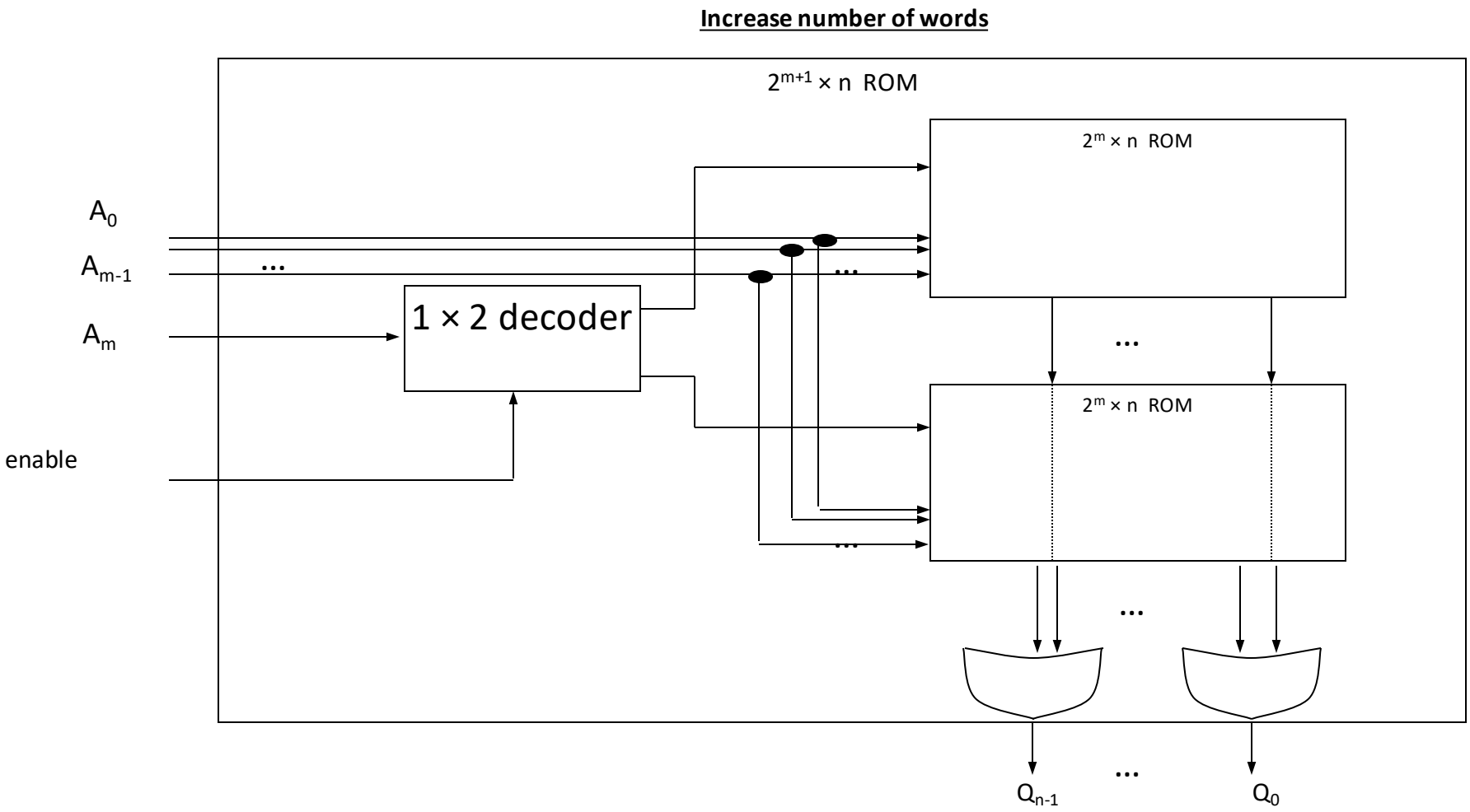
- Memory size needed often differs from size of readily available memories.
- When available memory is larger, simply ignore unneeded high-order address bits and higher data lines.

Increase width of words



Composing memory

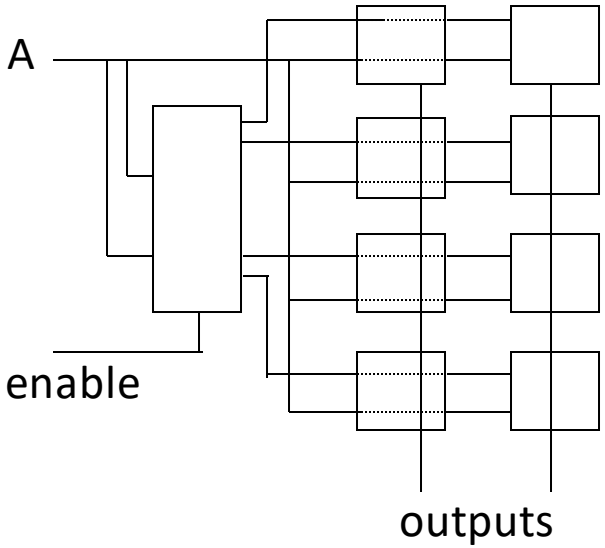
- Combine techniques to increase number and width of words



Composing memory

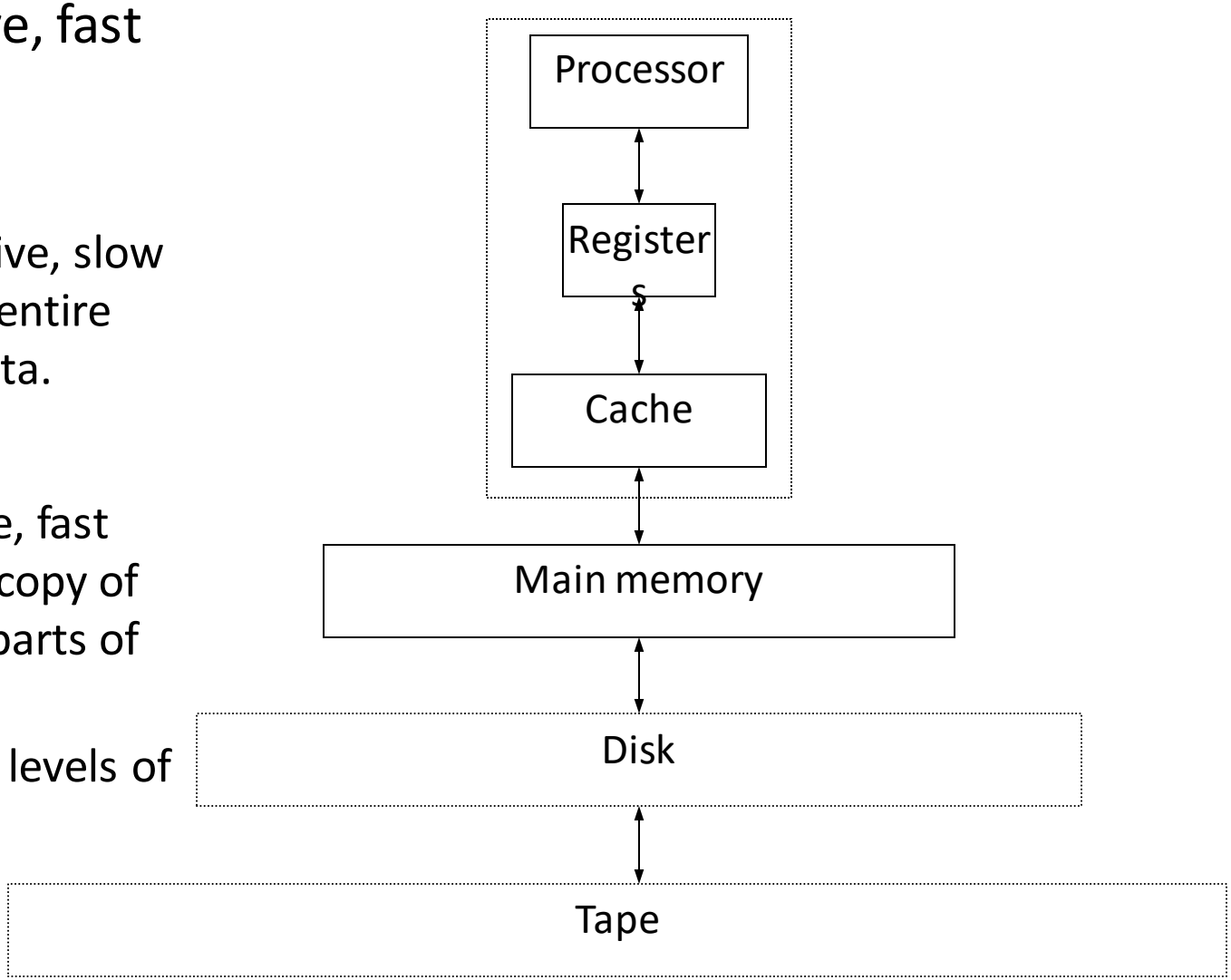
- When available memory is smaller, compose several smaller memories into one larger memory.
  - Connect side-by-side to increase width of words
  - Connect top to bottom to increase number of words
    - added high-order address line selects smaller memory containing desired word using a decoder.

Increase number and width of words



Memory hierarchy

- Want inexpensive, fast memory
- Main memory
  - Large, inexpensive, slow memory stores entire program and data.
- Cache
  - Small, expensive, fast memory stores copy of likely accessed parts of larger memory.
  - Can be multiple levels of cache.



Cache

- **Usually designed with SRAM**
  - faster but more expensive than DRAM
- **Usually on same chip as processor**
  - space limited, so much smaller than off-chip main memory
  - faster access (1 cycle vs. several cycles for main memory)
- **Cache operation:**
  - Request for main memory access (read or write)
  - First, check cache for copy
    - cache hit

      copy is in cache, quick access
    - cache miss

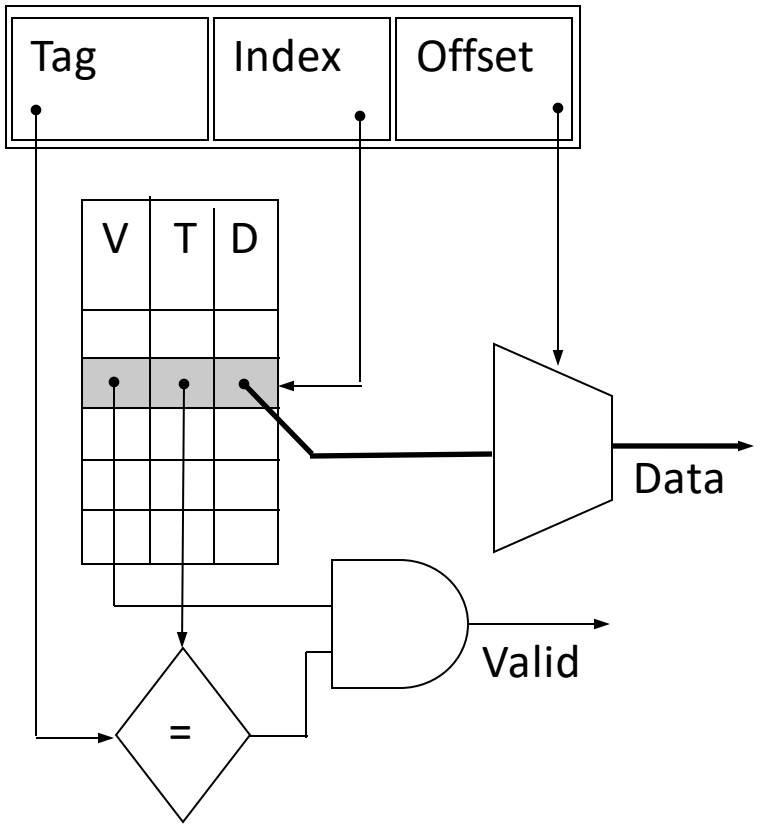
      copy not in cache, read address and possibly its neighbors into cache
- **Several cache design choices**
  - cache mapping, replacement policies, and write techniques

Cache mapping

- Far fewer number of available cache addresses
- Are address' contents in cache?
- Cache mapping used to assign main memory address to cache address and determine hit or miss.
- Three basic techniques:
  - Direct mapping
  - Fully associative mapping
  - Set-associative mapping
- Caches partitioned into indivisible blocks or lines of adjacent memory addresses
  - usually 4 or 8 addresses per line

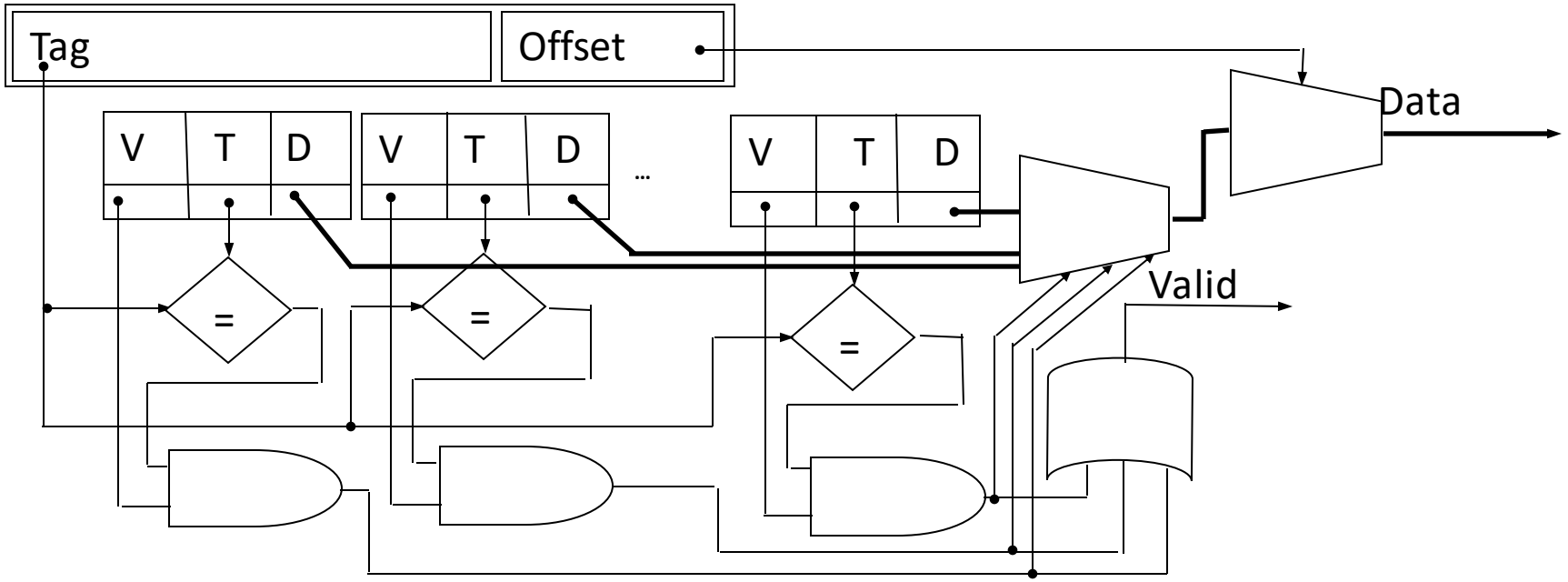
Direct mapping

- Main memory address divided into 2 fields
  - Index
    - cache address
    - number of bits determined by cache size
  - Tag
    - compared with tag stored in cache at address indicated by index
    - if tags match, check valid bit
- Valid bit
  - indicates whether data in slot has been loaded from memory
- Offset
  - used to find particular word in cache line



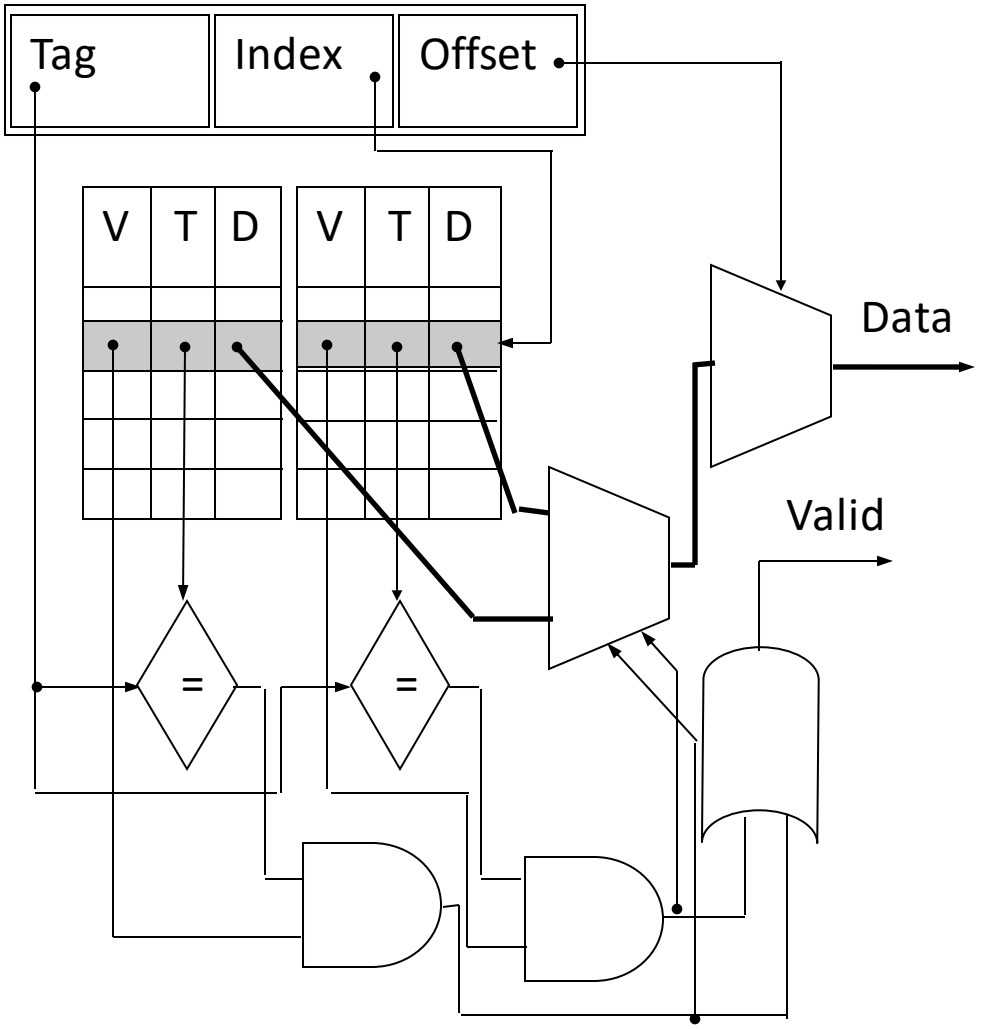
Fully associative mapping

- Complete main memory address stored in each cache address.
- All addresses stored in cache simultaneously compared with desired address.
- Valid bit and offset same as direct mapping.



Set-associative mapping

- Compromise between direct mapping and fully associative mapping.
- Index same as in direct mapping.
- But, each cache address contains content and tags of 2 or more memory address locations.
- Tags of that set simultaneously compared as in fully associative mapping.
- Cache with set size N called N-way set-associative
  - 2-way, 4-way, 8-way are common



Cache-replacement policy

- Technique for choosing which block to replace
  - when fully associative cache is full.
  - when set-associative cache's line is full.
- Direct mapped cache has no choice
- Random
  - replace block chosen at random.
- LRU: least-recently used
  - replace block not accessed for longest time.
- FIFO: first-in-first-out
  - push block onto queue when accessed.
  - choose block to replace by popping queue.

Cache write techniques

When written, data cache must update main memory

- Write-through
  - write to main memory whenever cache is written to
  - easiest to implement
  - processor must wait for slower main memory write
  - potential for unnecessary writes
- Write-back
  - main memory only written when “dirty” block replaced
  - extra dirty bit for each block set when cache block written to
  - reduces number of slow main memory writes

Cache impact on system performance

- Most important parameters in terms of performance:
  - Total size of cache
    - total number of data bytes cache can hold
    - tag, valid and other house keeping bits not included in total
  - Degree of associativity
  - Data block size
- Larger caches achieve lower miss rates but higher access cost
  - e.g.,
    - 2 Kbyte cache: miss rate = 15%, hit cost = 2 cycles, miss cost = 20 cycles
      avg. cost of memory access = (0.85 * 2) + (0.15 * 20) = 4.7 cycles
    - 4 Kbyte cache: miss rate = 6.5%, hit cost = 3 cycles, miss cost will not change
      avg. cost of memory access = (0.935 * 3) + (0.065 * 20) = 4.105 cycles
      (improvement)
    - 8 Kbyte cache: miss rate = 5.565%, hit cost = 4 cycles, miss cost will not change
      avg. cost of memory access = (0.94435 * 4) + (0.05565 * 20) = 4.8904 cycles
      (worse)

Cache performance trade-offs

- Improving cache hit rate without increasing size
  - Increase line size
  - Change set-associativity