ELE 455/555
Computer System Engineering

Section 1 – Review and Foundations

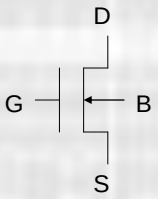
Class 3 – Technology

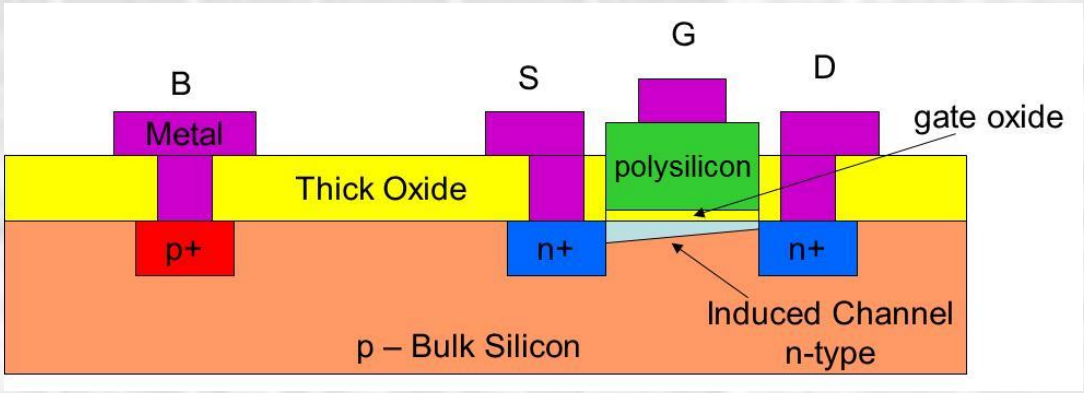
Technology
MOSFETs

- MOSFET Terminology
  - Metal Oxide Semiconductor Field Effect Transistor
  - 4 terminal device
    - Source, Gate, Drain, Body
  - Threshold Voltage (Vth or Vt)
    - The voltage from gate to source (Vgs) required to “turn on” the device
  - Ron / Roff
    - Device impedance in the “on” or “off” state
  - Leakage Current
    - Parasitic current from junctions to substrate or gate to junctions

Technology
MOSFETs

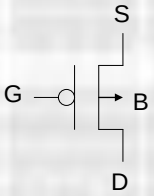
- MOSFET Terminology
  - N-type (N-channel)
    - Forms a conducting n-channel from source to drain
    - Requires a positive Vgs > Vth to form the channel and “turn on” the device

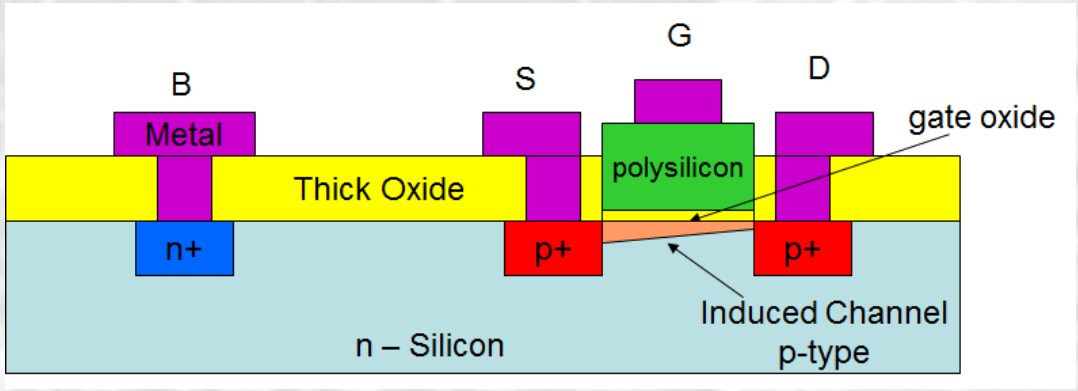




Technology
MOSFETs

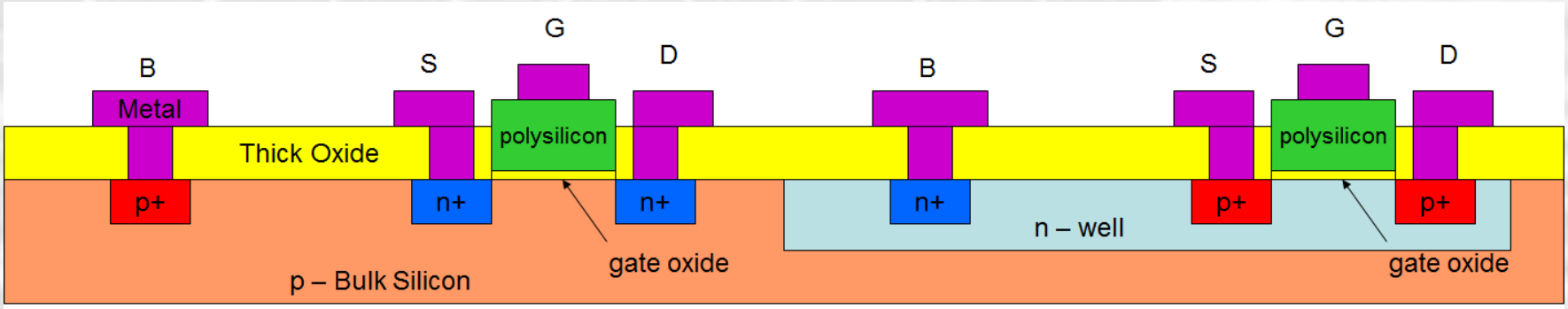
- MOSFET Terminology
  - P-type (P-channel)
    - Forms a conducting p-channel from source to drain
    - Requires a negative Vgs > Vth to form the channel and “turn on” the device





Technology
MOSFETs

- MOSFET Terminology
  - CMOS
    - Complementary MOS
    - Contains both P-MOS and N-MOS devices
    - Almost all digital circuits today are built using CMOS technology

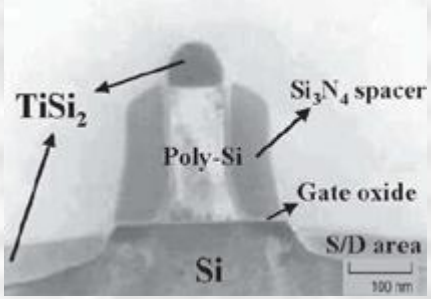


N-MOS P-MOS

Technology

CMOS Technology Trends

- CMOS Process Technology
  - Traditionally referenced to the gate length or metal 1/2 pitch
    - 0.25 micron process, 130nm process
  - Over the last 5 years the relationship between reference name and physical parameters has become tenuous
    - Current generation processes are 22nm shifting to 14nm
    - 22nm = 22×10^{-9} meters = 22×10^{-6} millimeters = 22 millionths of a mm
    - 1 silicon atom ~ 5 angstrom spacing = 5×10^{-10}m
      → 22nm = 44 atoms
  - Gate oxide thickness is 2nm
    → 4 rows of atoms

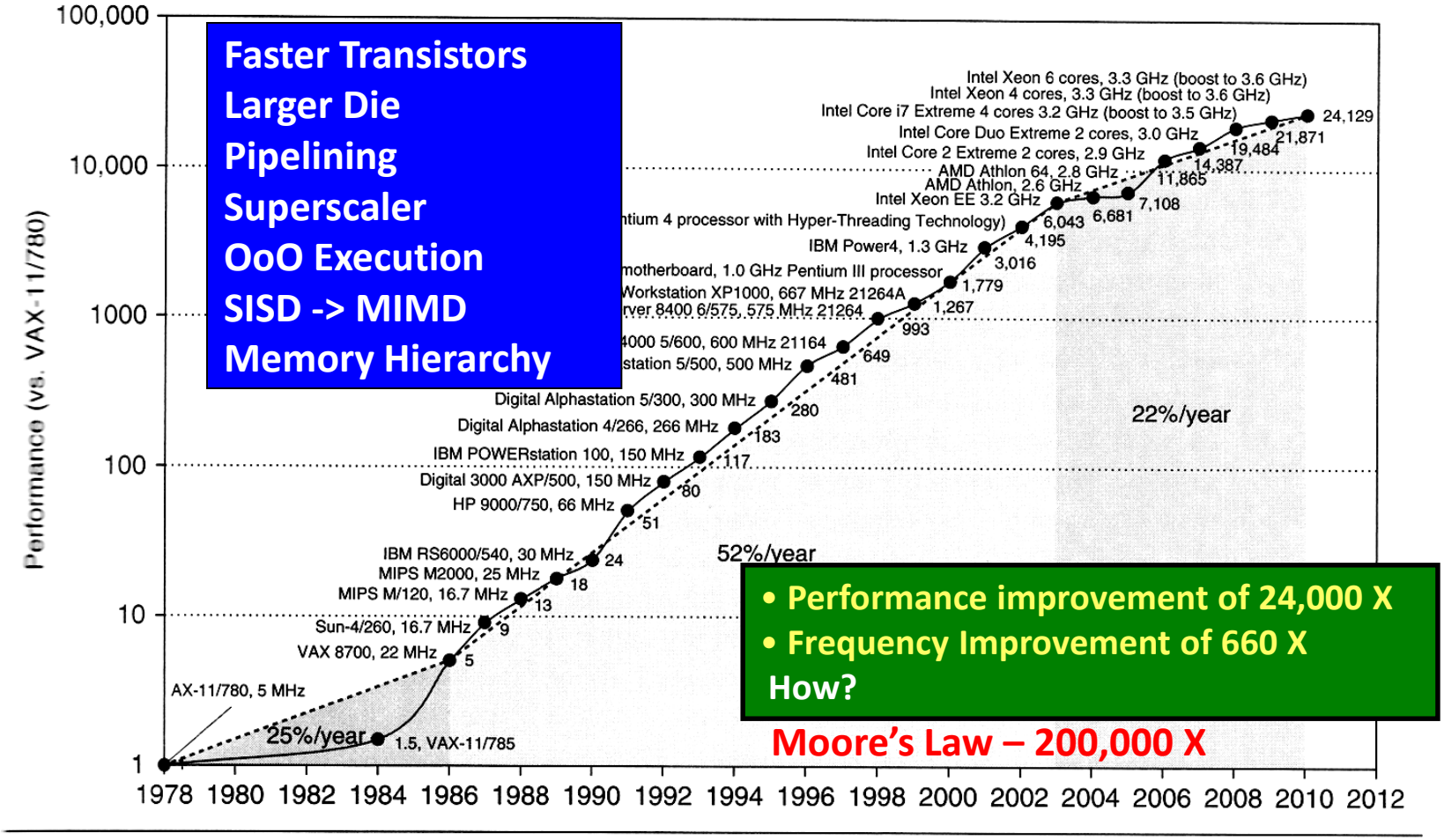


Src: TMS - Vol 57, No. 9

Technology

CMOS Technology Trends

- Processor Performance

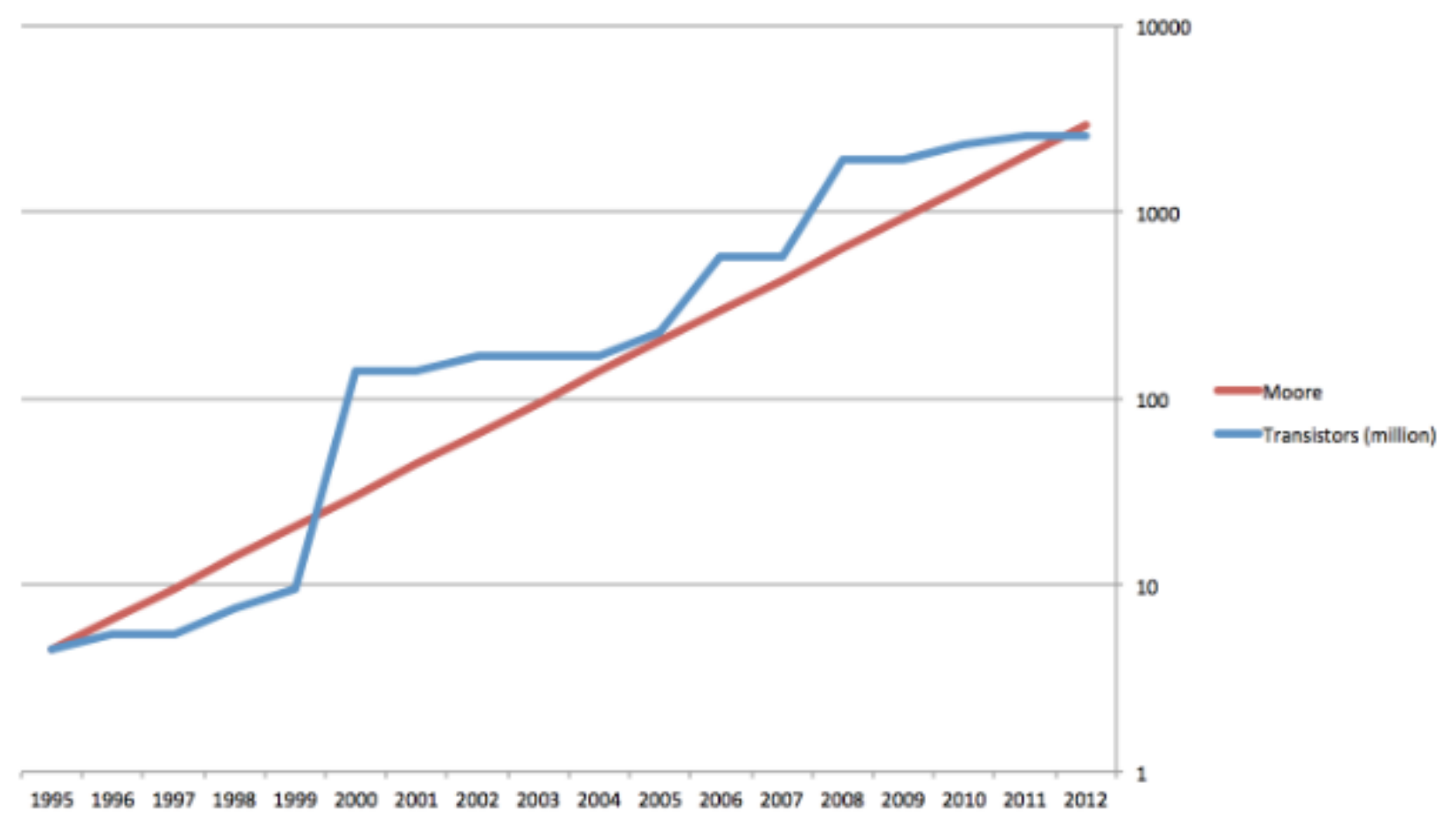


Source: Computer Architecture, Hennessy and Patterson, 2012 Elsivier Inc

Technology

CMOS Technology Trends

- Transistors per chip

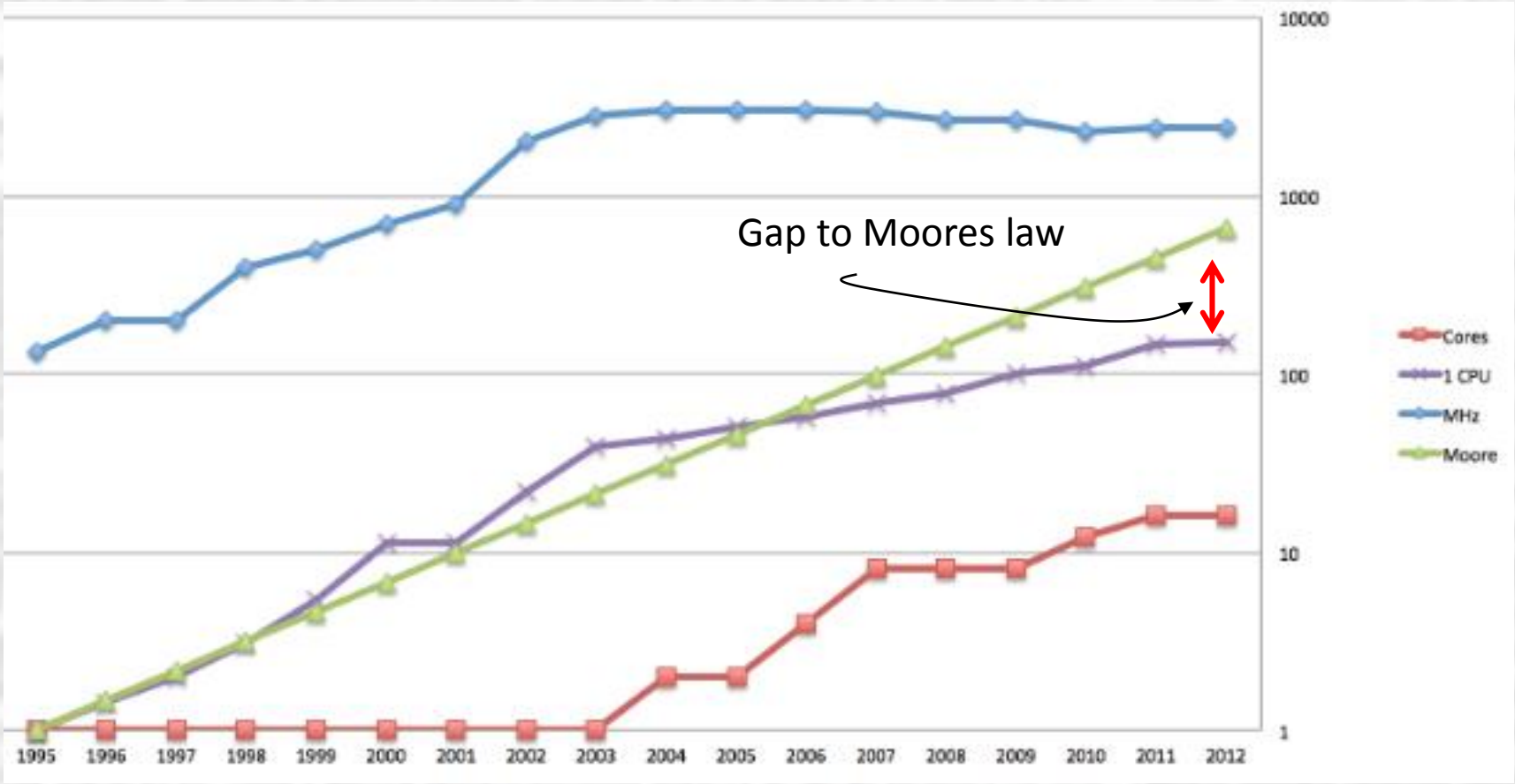


src: http://www.saphana.com

Technology

CMOS Technology Trends

- SPECint Performance – single CPU

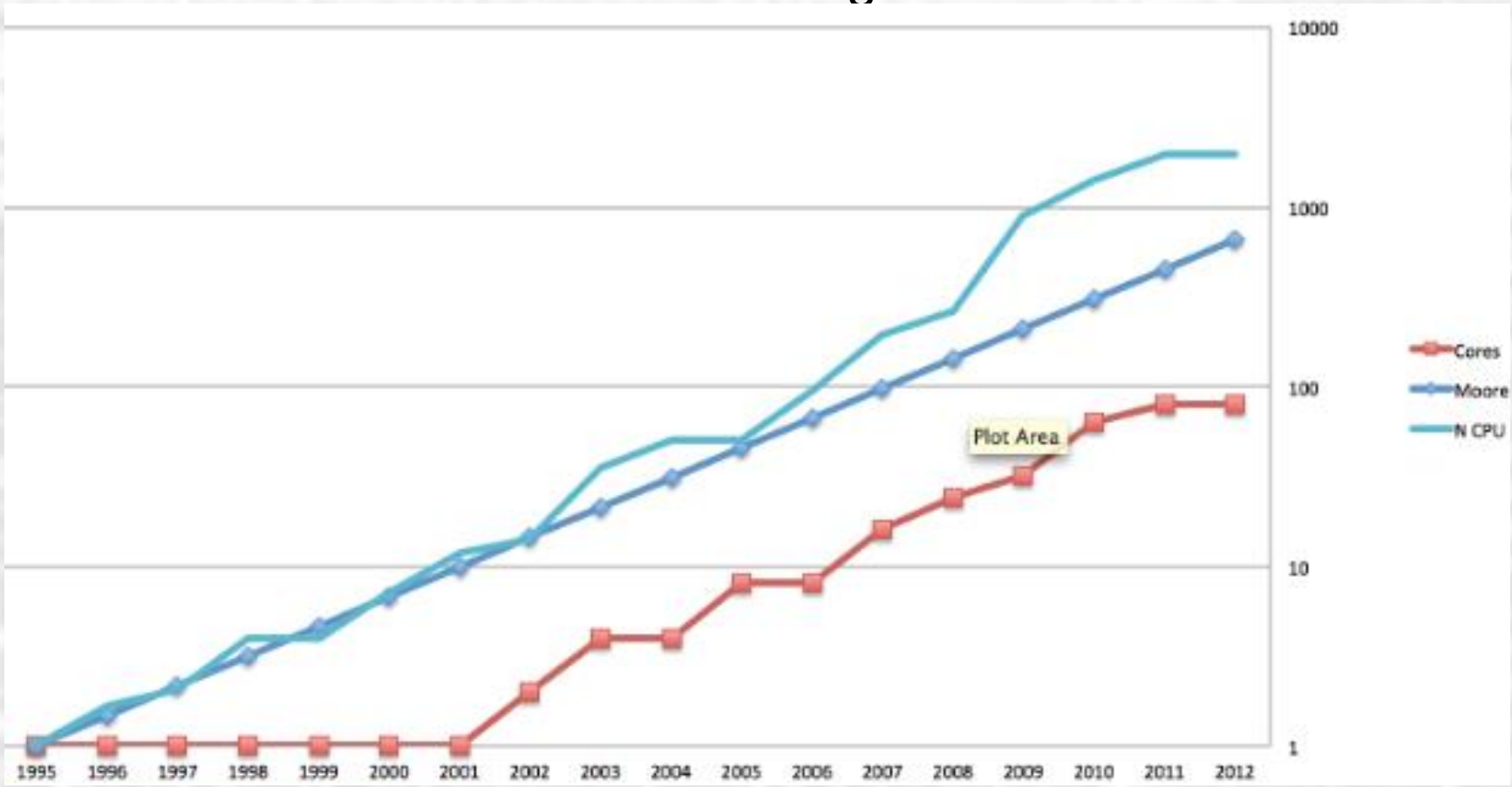


src: http://www.saphana.com

Technology

CMOS Technology Trends

- SPECint Performance – Including multi-core

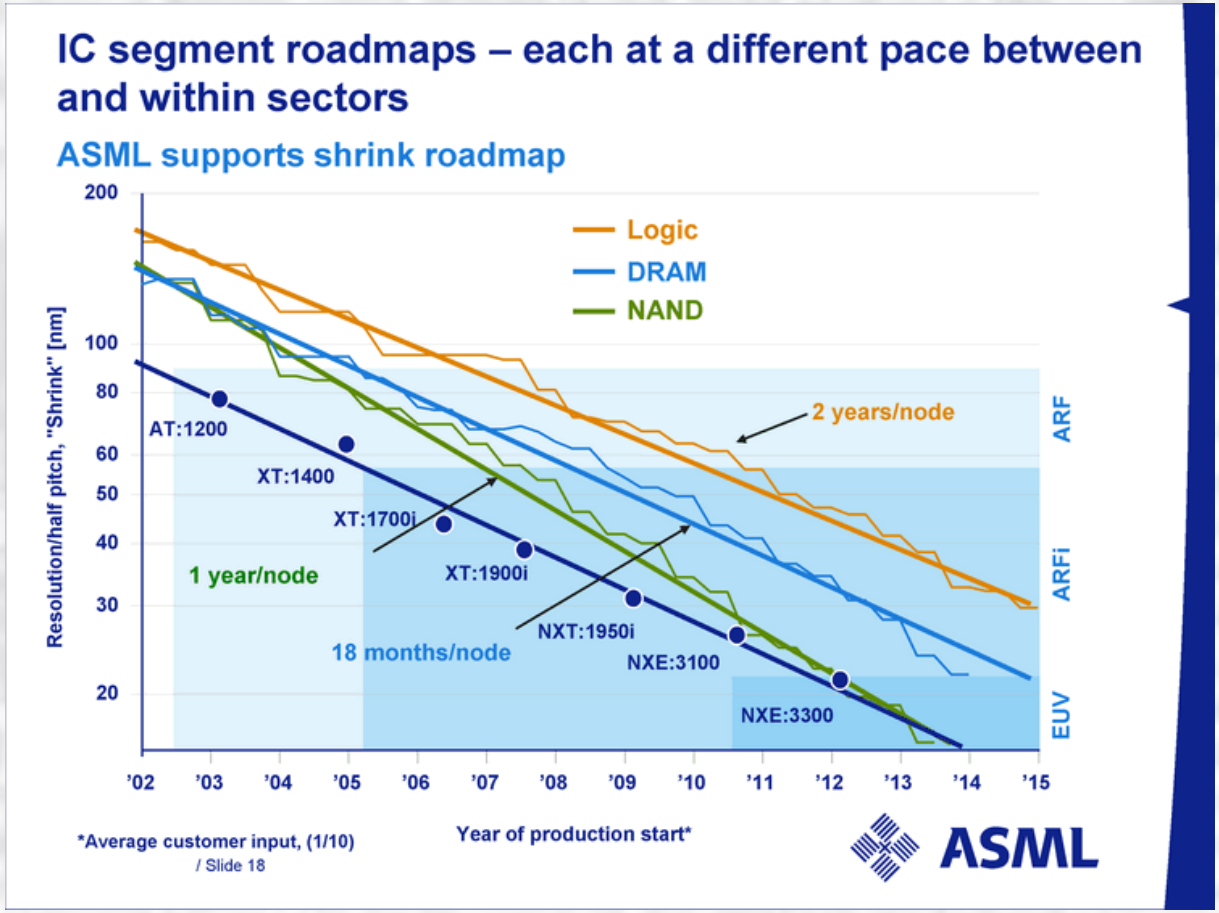


src: http://www.saphana.com

Technology

CMOS Technology Trends

- Half-Pitch Trends



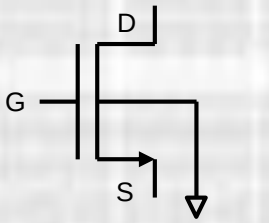
Technology
CMOS Logic

- Simplified CMOS Digital Design
  - Use simplified models for the MOSFETs
    - switch – either “on” or “off”

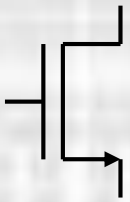
  - N-MOS devices
    - “on” when a logic high is applied to the gate
    - “off” when a logic low is applied to the gate
  - P-MOS devices
    - “on” when a logic low is applied to the gate
    - “off” when a logic high is applied to the gate

Technology
CMOS Logic

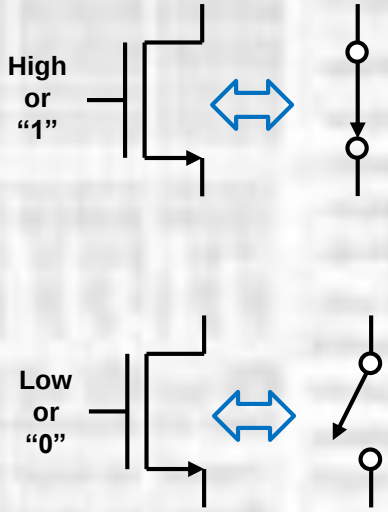
- Simplified CMOS Digital Design
  - N-MOS devices
    - "on" when a logic high is applied to the gate
    - "off" when a logic low is applied to the gate



Vss or GND

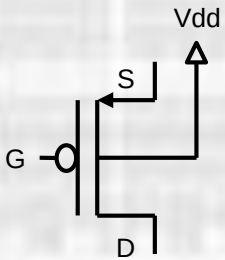


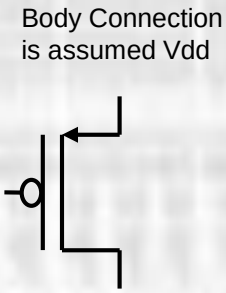
Body Connection
is assumed
Vss or GND

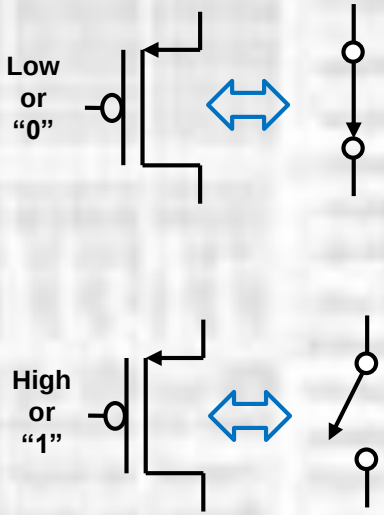


Technology
CMOS Logic

- Simplified CMOS Digital Design
  - P-MOS devices
    - "on" when a logic low is applied to the gate
    - "off" when a logic high is applied to the gate

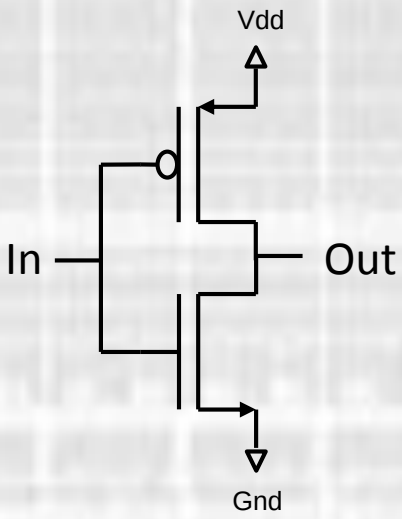


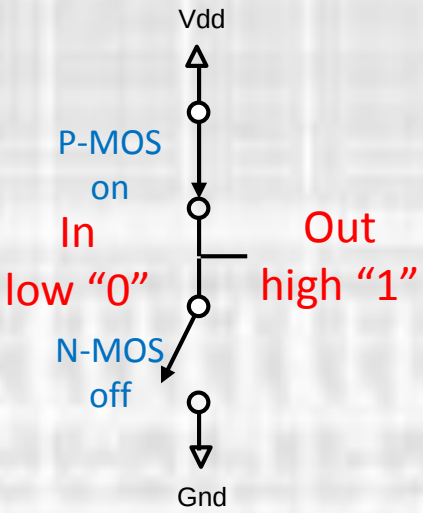


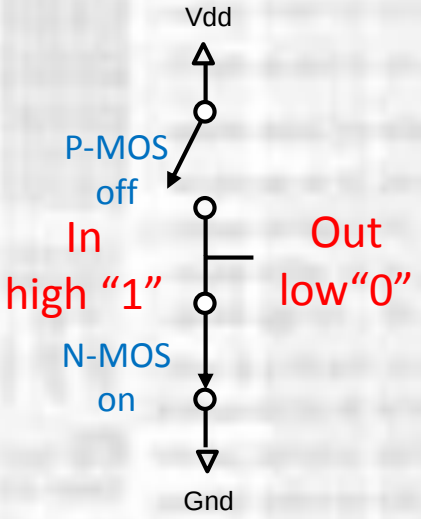


Technology
CMOS Logic

- Simplified CMOS Digital Design
  - CMOS Inverter
    - logic low “0” is Gnd
    - logic high “1” is Vdd



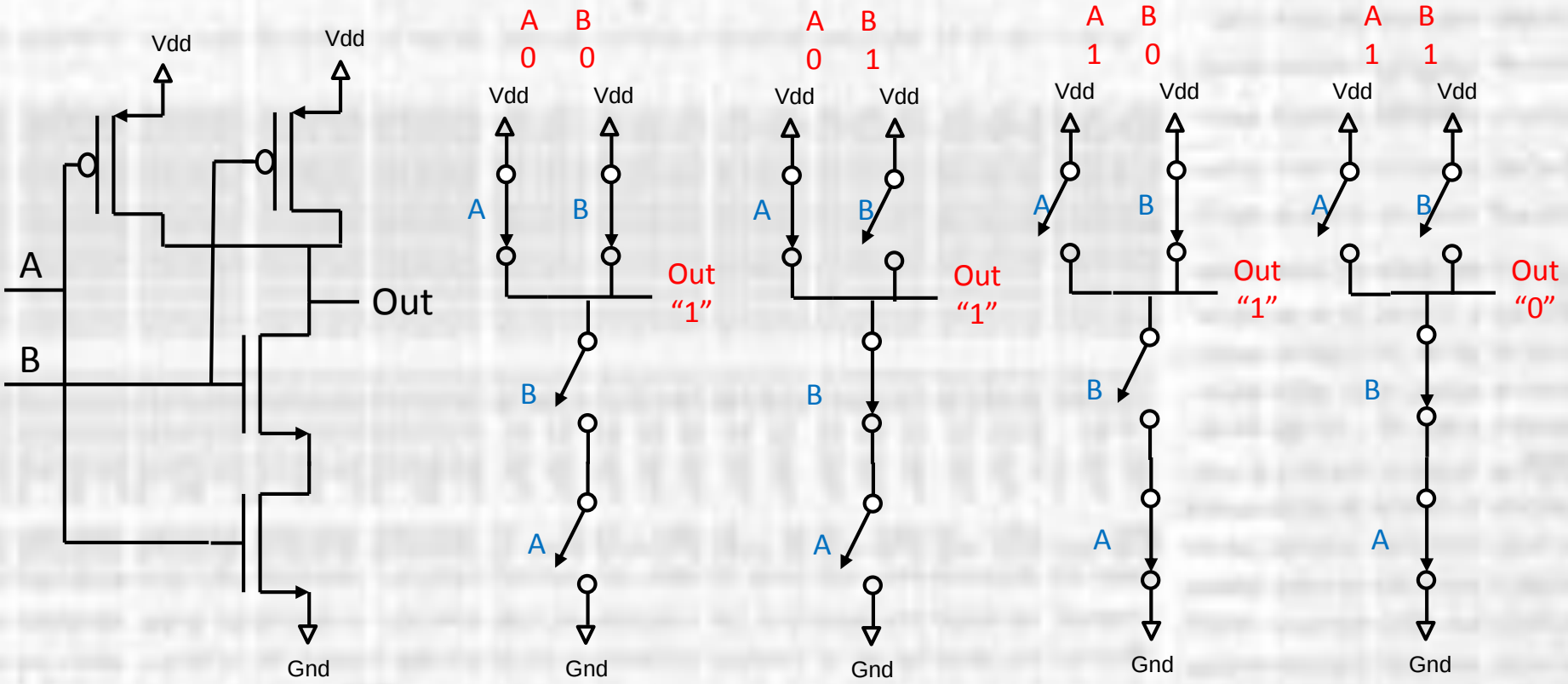




Technology
CMOS Logic

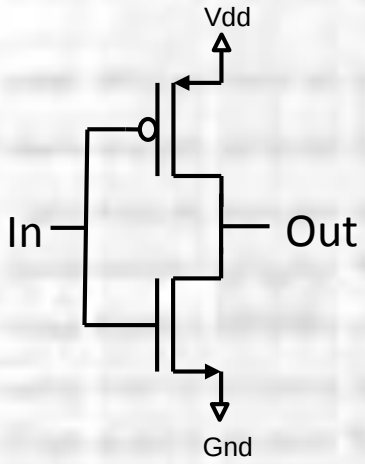
- Simplified CMOS Digital Design
  - CMOS Nand Gate

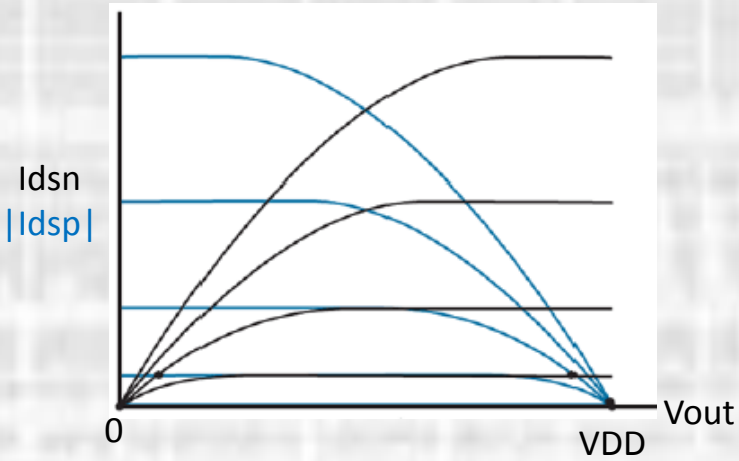
A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0

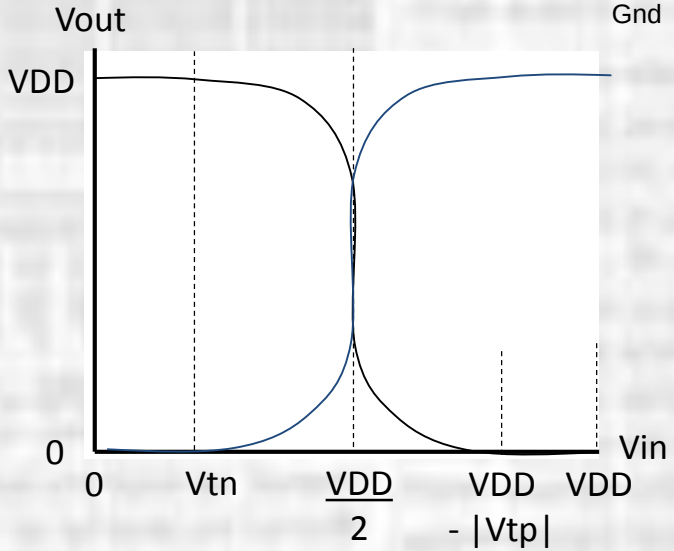


Technology
CMOS Logic

- Gate Level Performance
  - DC Characteristics
    - $I_{DS} = \beta[(V_{GS} - V_t)V_{DS} - {V_{DS}}^2/2]$ - linear region
    - $I_{DS} = \beta/2\,(V_{GS} - V_t)^2$ - saturated region
    - $\beta = \frac{\mu\varepsilon}{t_{ox}}\frac{W}{L}$ $\beta_n \cong 2.8\beta_p$

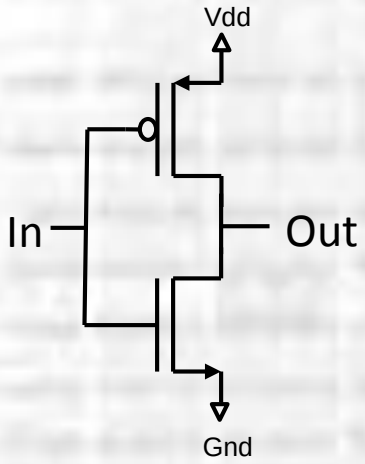


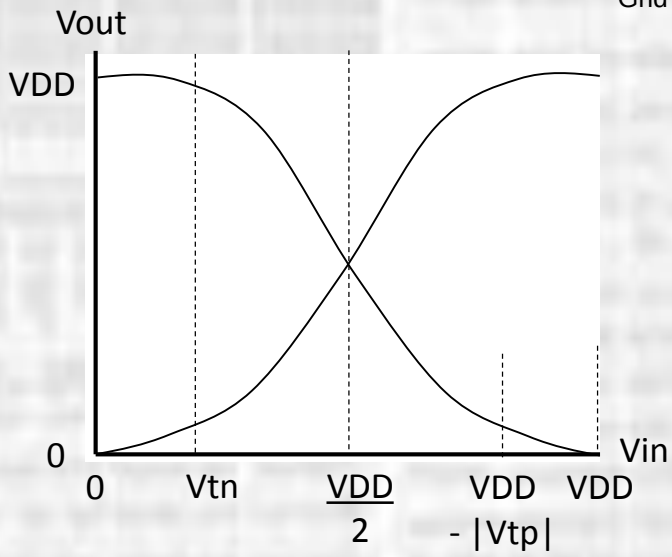




Technology
CMOS Logic

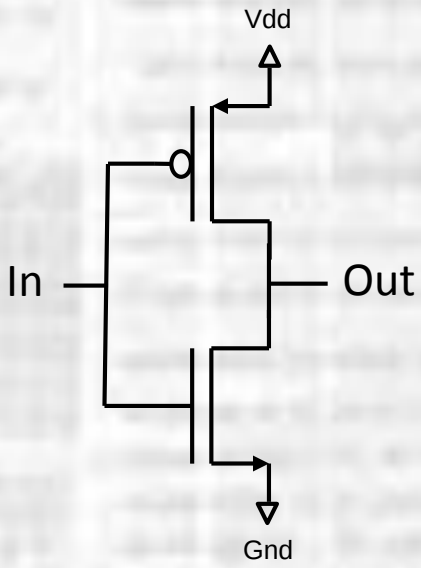
- Gate Level Performance
  - Transient Characteristics
    - Equalize t_r and t_f by making Wp ≅ 2.8 Wn
    - $t_r \cong t_f \cong K_n \frac{C_L}{\beta_n V_{DD}} \cong K_p \frac{C_L}{\beta_p V_{DD}}$, $Kn \sim Kp \sim 3.5$
    - $t_d \cong t_r/2 \cong t_f/2$
    - Optimize delay
      - reduce C_L
      - increase V_{DD}
      - increase β → reduce t_{ox}
        - why might this be the best approach?





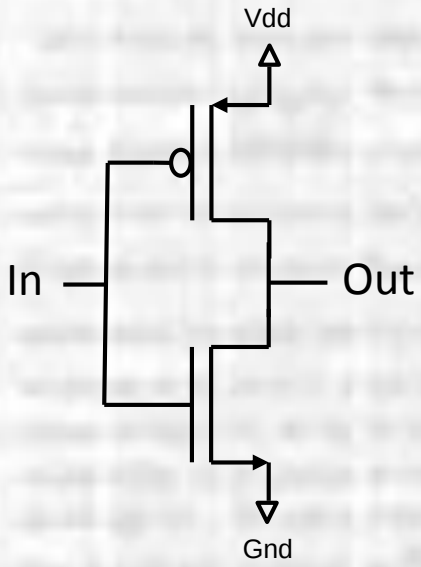
Technology
CMOS Logic

- Gate Level Power
  - 3 primary components of gate level power
  - Static Power (leakage)
  - Dynamic Power (CV^2F)
  - Short Circuit Power (shoot-through)



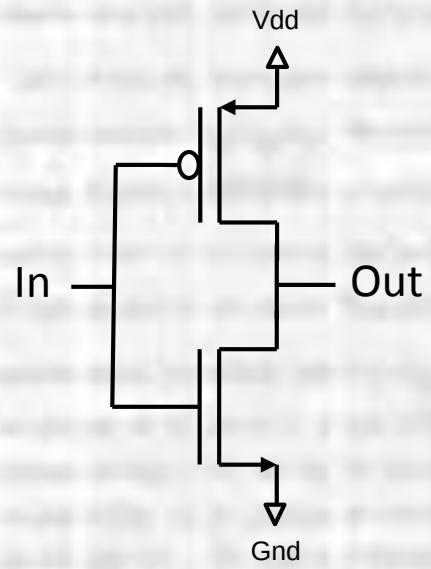
Technology
CMOS Logic

- Gate Level Power
  - Static Power
    - Leakage currents through the reverse biased diode junctions – always present
    - Sub-threshold current – current from S-D when the input voltage is below Vt – due to voltage drops
    - Gate leakage current – current from the gate to S/D/Body – due to oxide defects or quantum tunneling
  - Design Considerations
    - Multiple Vt devices
    - Thick oxide devices
    - Reduce supply voltages – why?



Technology
CMOS Logic

- Gate level Power
  - Dynamic Power
    - Power associated with slewing the load capacitance
    - $E = \int Pdt = \int IVdt = \int C\,{}^{dV}/_{dt}\,Vdt = \int CVdv = \frac{CV^2}{2}$
    - energy per transition = $C_L V_{dd}^2/2$
    - Power is energy / time
    - $P_{dynamic} = CV^2f$
  - Design considerations
    - Run circuits at the lowest possible speed
    - Reduce supply voltages
    - Minimize capacitance

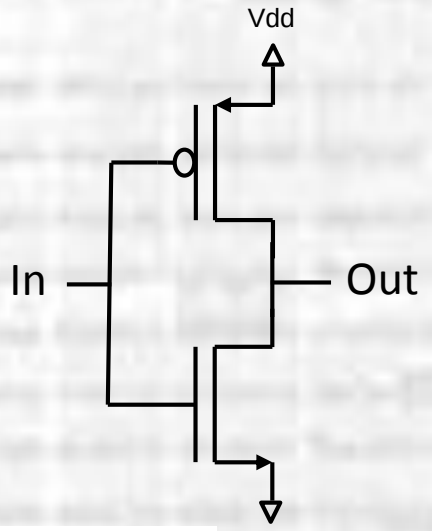


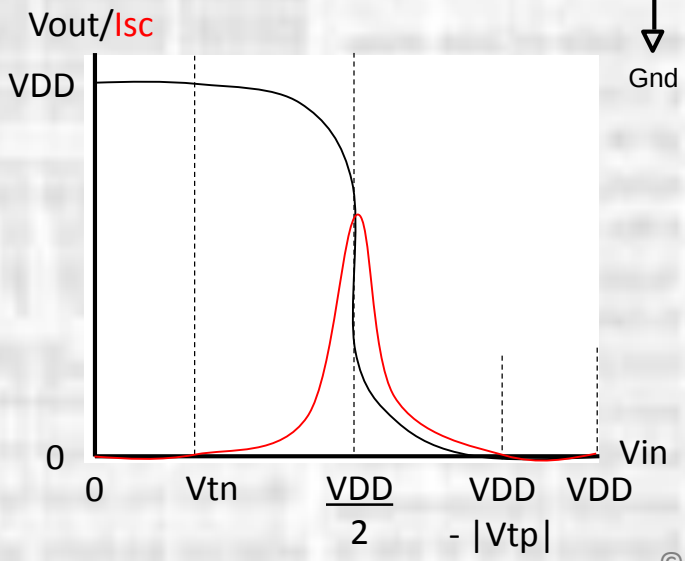
Technology
CMOS Logic

- Gate Level Power
  - Short Circuit Power
    - Both devices are on
    - Symmetrical gate →

$$P_{sc} \cong \frac{\beta}{12}(V_{DD} - 2V_t)^3 f t_{rf}$$

  - Design Considerations
    - large relative loads → lower Psc
    - small relative loads → higher Psc
    - Match t_r and t_f through the chain
    - Reduce supply voltages





Technology

CMOS Integrated Circuits

- Chip Level Considerations
  - Millions to hundreds of millions of gates
    - Physical Space
    - Sheer numbers of gates
    - Keeping shared resources close to multiple users (memory)
    - I/O pin access and placement
    - Interconnect – getting all the wires connected
      - Typical processes have 6 – 10 layers of interconnect
      - Cell, local, global, power
    - Performance
    - Power / Heat Dissipation

Technology

CMOS Integrated Circuits

- Chip Level Considerations
  - Performance Drivers
    - Process Technology
      - Transistor performance
      - Short channel vs. long channel devices
      - High Vt and Low Vt devices
    - Clock Frequency
      - Maximum is set by the longest unit delay
      - Very complex timing tools used to ensure max frequency
    - Interconnect
      - RC delays
      - Capacitive coupling

Technology
CMOS Integrated Circuits

- Chip Level Considerations
  - Power Drivers
    - Process Technology - Dynamic, Static, Short circuit (D/S/SC)
    - Number of gates – D/S/SC
    - Clock Frequency – D/SC
      - Dynamic power becomes CV^2f, where f is clock frequency
      - Short circuit power is also multiplied by f
    - Supply Voltage – D/S/SC
    - Routing Efficiency – D/SC
      - Minimizing capacitance is critical

Technology

CMOS Integrated Circuits

- Chip Level Considerations
  - Power / Performance Balance
    - Device Level Solutions
      - Multiple Vth devices
        - Low Vth devices for high performance paths
        - High Vth devices for low performance paths
    - Chip Level Solutions
      - Reduced interconnect R and C
      - Power islands – gating the power to circuits not in use
      - Clock Gating – turn off the clocks to circuits not in use

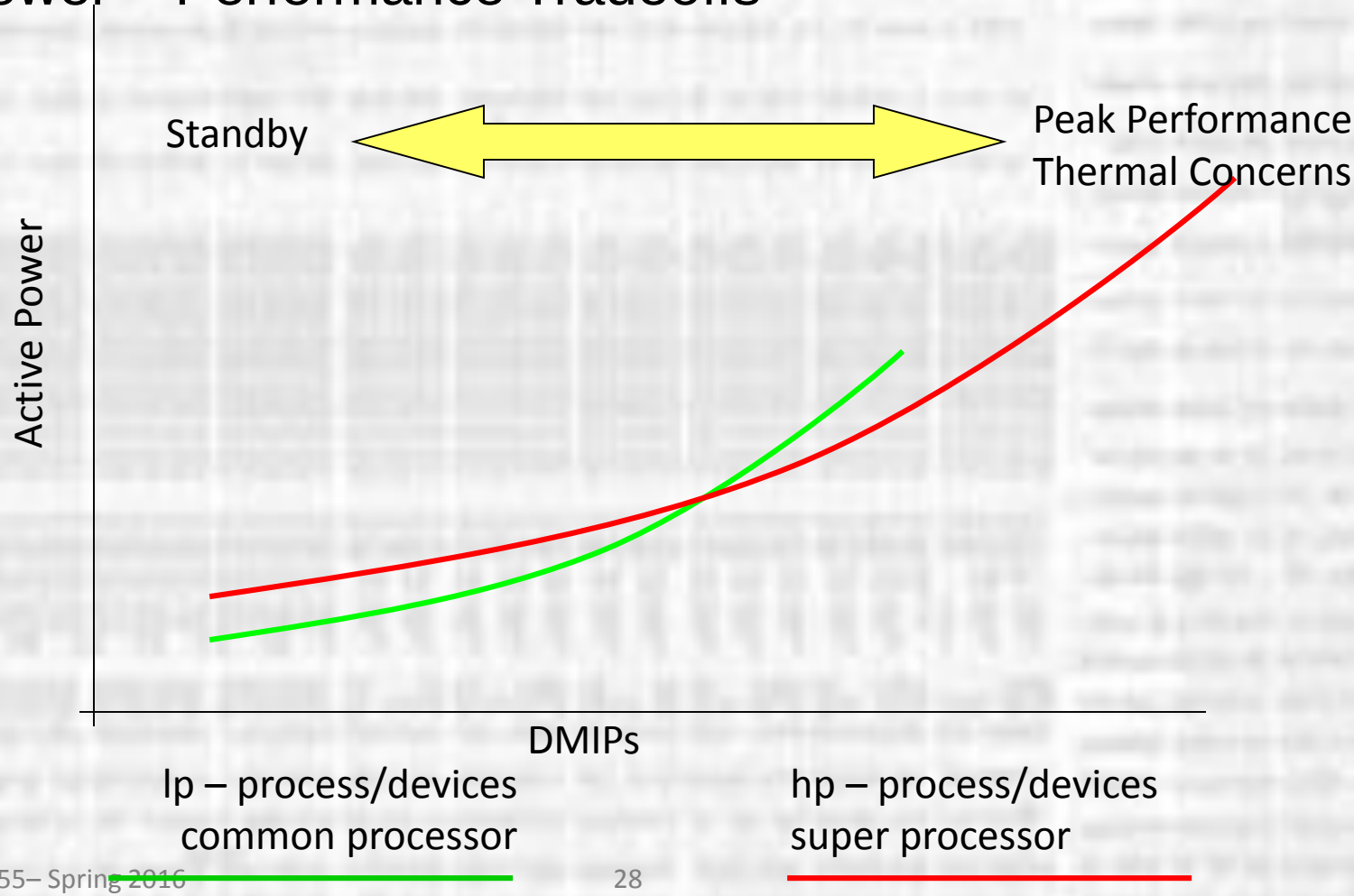
Technology
CMOS Integrated Circuits

- Chip Level Considerations
  - Power / Performance Balance
    - System Level Solutions
      - Dynamic Voltage Scaling – changing VDD as needed
      - Dynamic Frequency Scaling – changing the clock frequency as needed
      - Together these are referred to as DVFS
    - Architectural Solutions
      - Pipelining
      - Multi-core processors
        - Homogeneous – dual/quad core
        - Heterogeneous – big/little
      - Memory Hierarchy

Technology

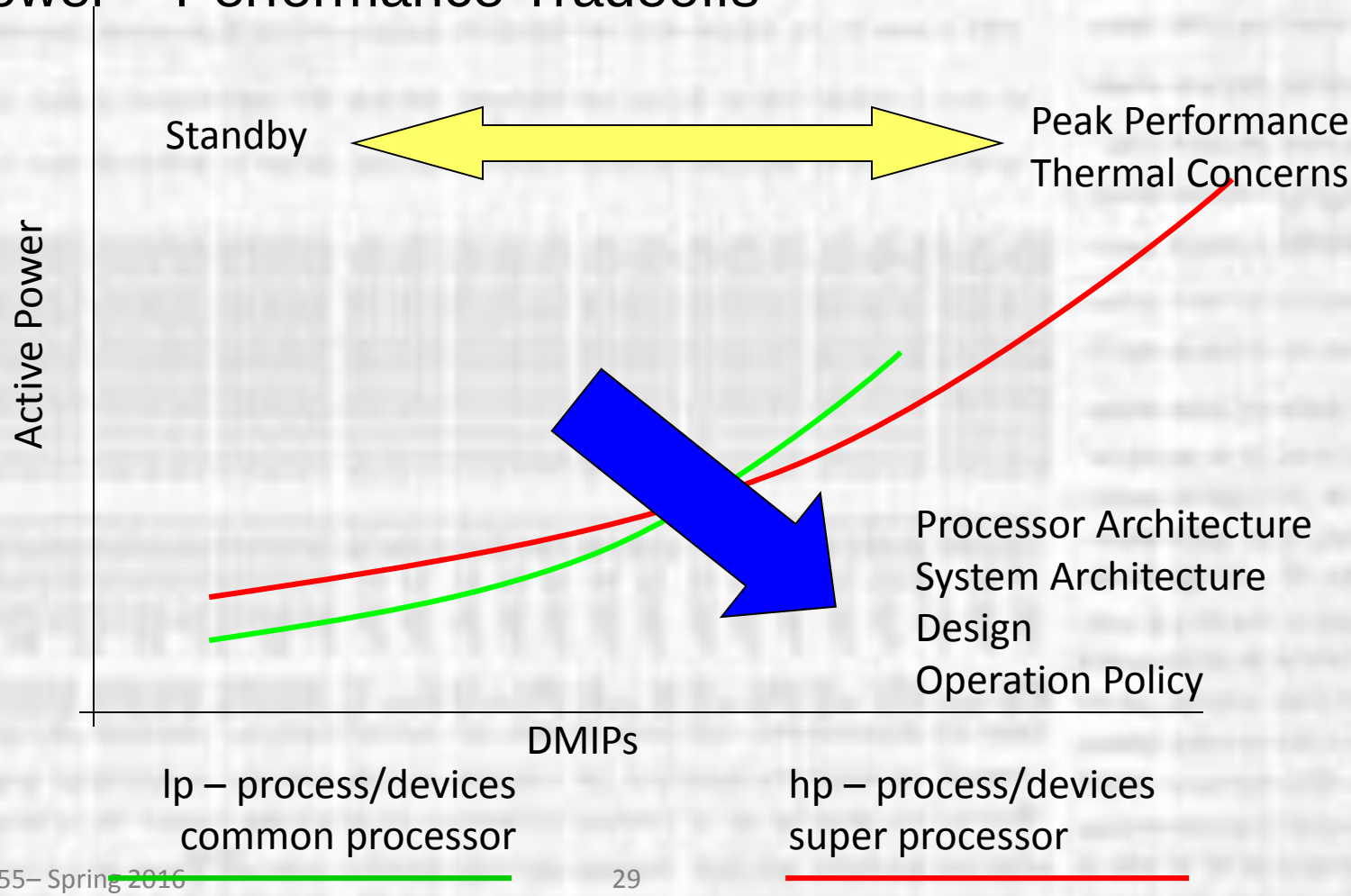
CMOS Integrated Circuits

- Power – Performance Tradeoffs



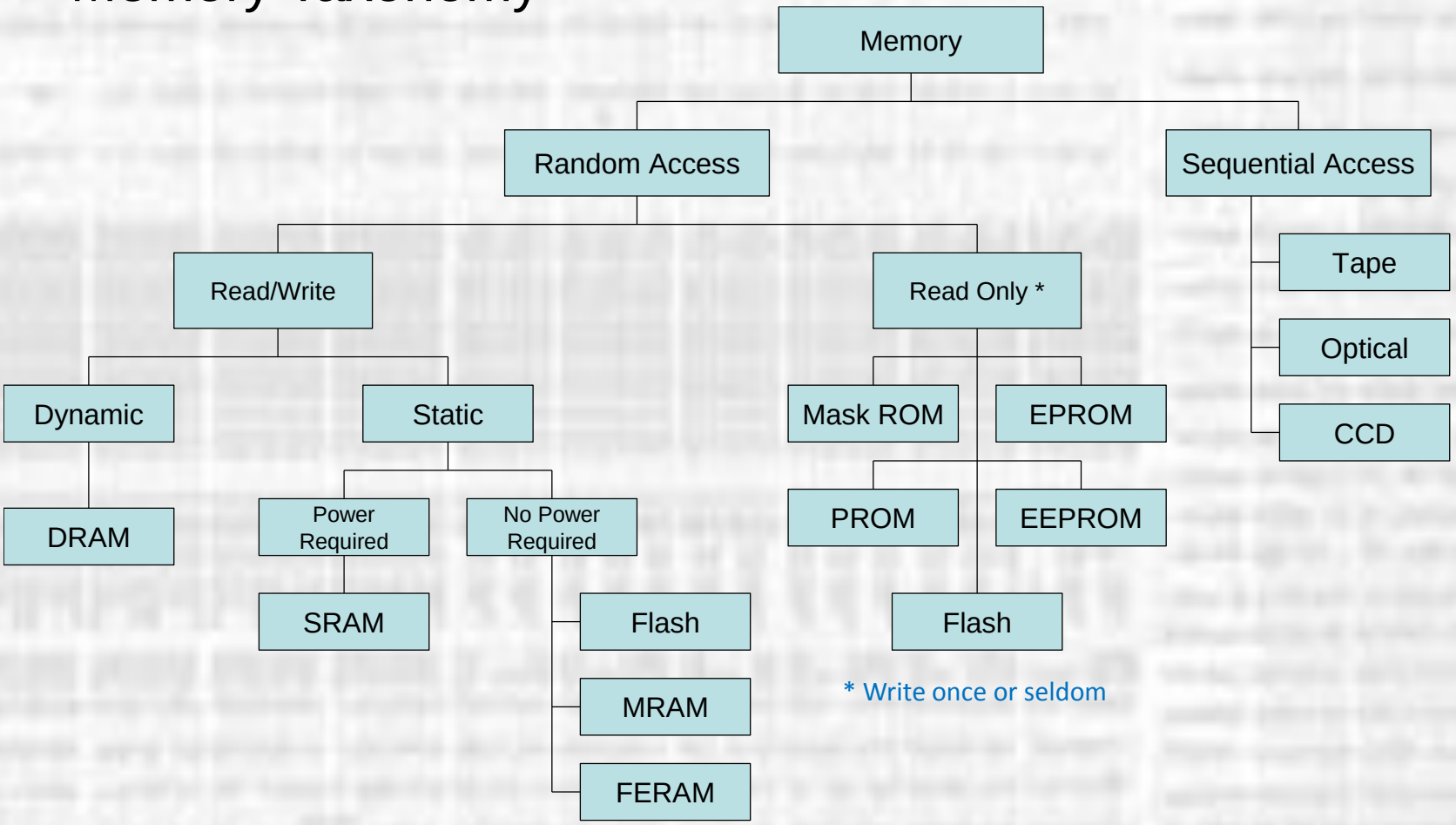
Technology
CMOS Integrated Circuits

- Power – Performance Tradeoffs



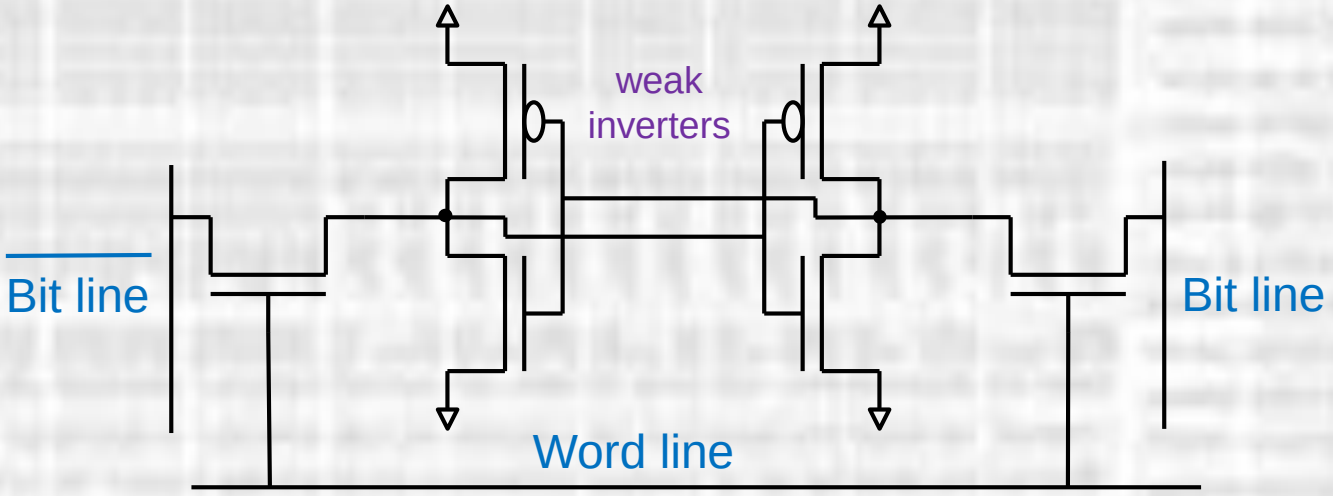
Technology
Memory

- Memory Taxonomy



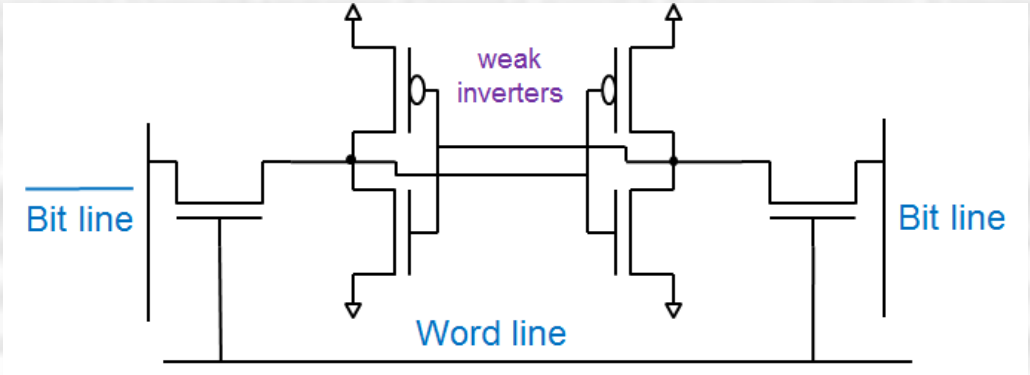
Technology
Memory

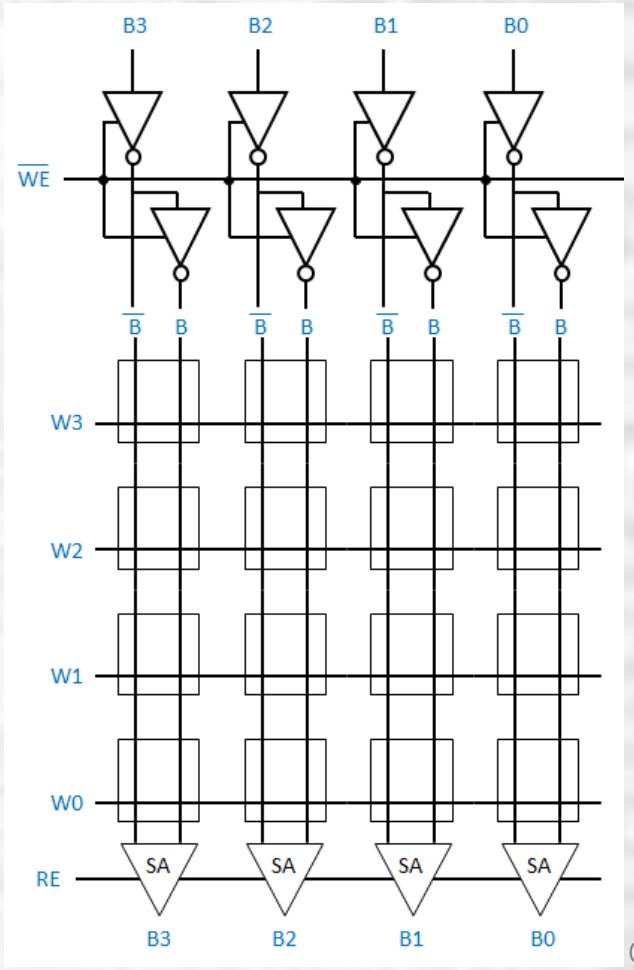
- SRAM – Static Random Access Memory
  - Memory cell (1 bit) is based on a feedback circuit
  - Bit value is retained as long as power is maintained
  - Fastest read/write (R/W)
  - Highest power
  - Lowest density
  - Used in caches and small data memories



Technology
Memory

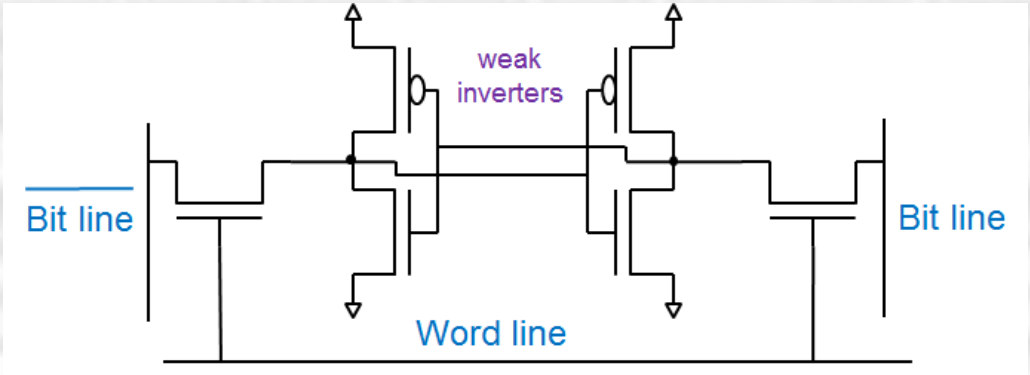
- SRAM – Static Random Access Memory
  - Write
    - All Word lines low
    - Read Enable (RE) disabled (low)
    - Place B0, B1, B2, B3 on inputs
    - Pull write enable bar ($\overline{WE}$) low
    - Strobe the desired word line high
    - Bit lines override the bit cell inverters and store the new value in the cell

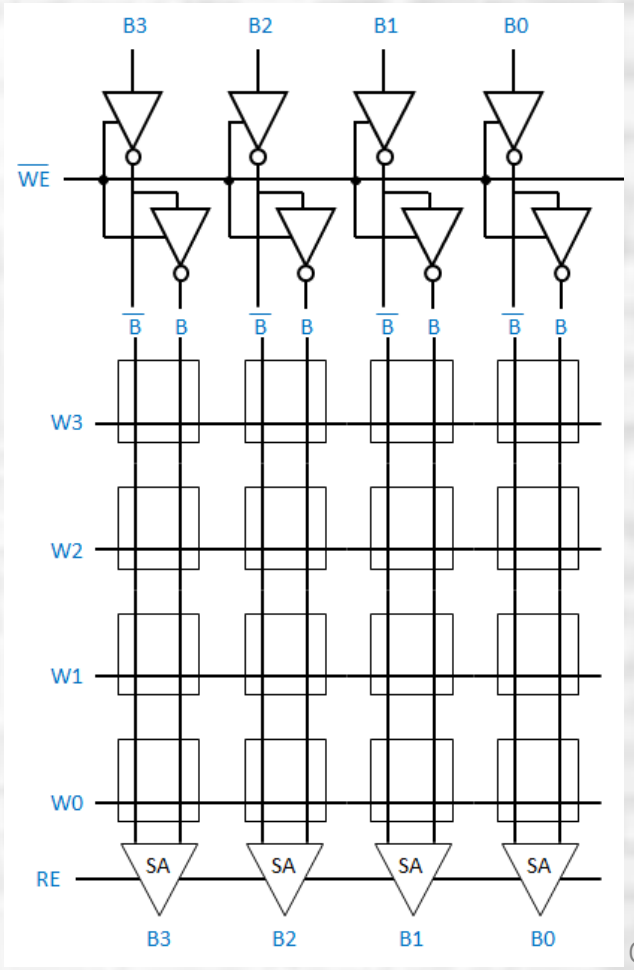




Technology
Memory

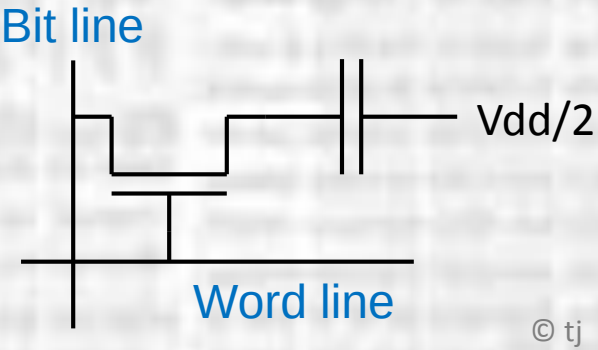
- SRAM – Static Random Access Memory
  - Read
    - All Word lines low
    - Write enable bar ($\overline{WE}$) high
      - inverters tristated
    - Read Enable (RE) high
    - Strobe the desired word line high
    - Bit cell inverters drive the bit lines and sense amplifiers read the value





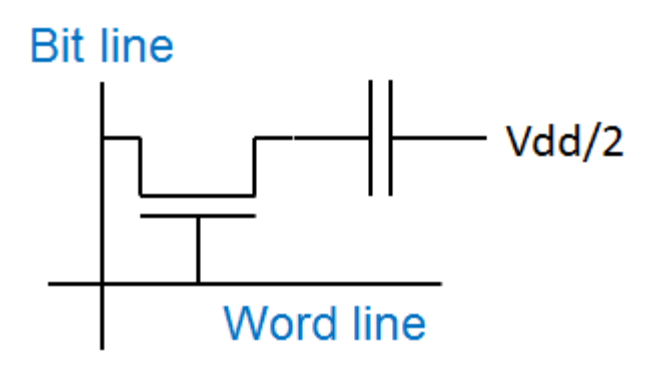
Technology
Memory

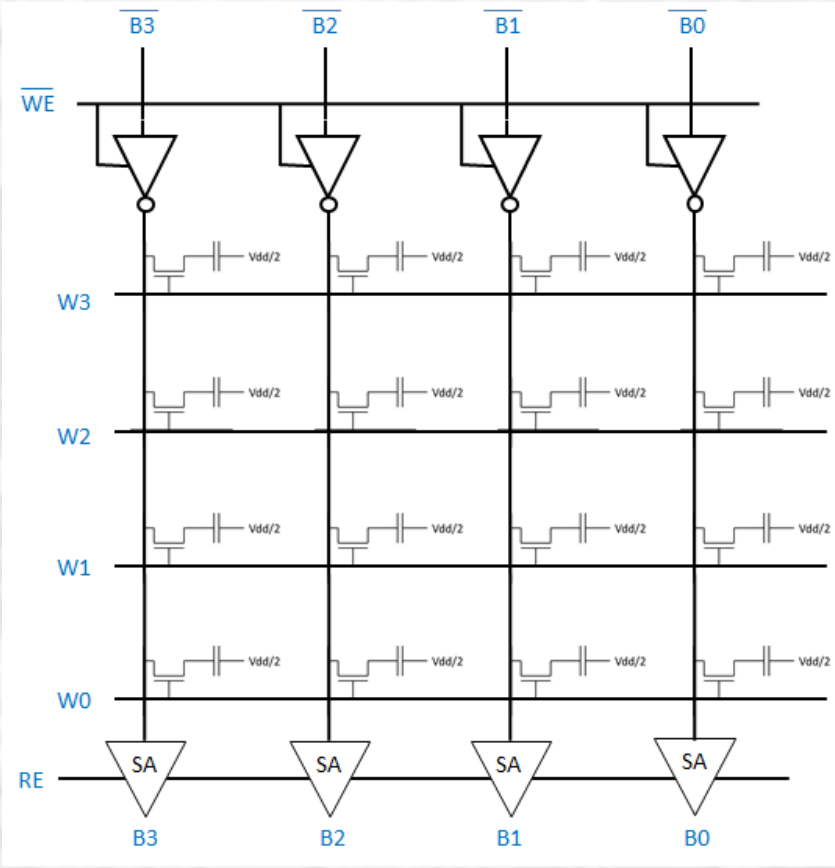
- SDRAM – Synchronous Dynamic Random Access Memory
  - Memory cell (1 bit) is based on capacitor charge storage
  - Bit value decays over time
    - must be recharged – called a refresh cycle
  - Standard SDRAM transfers 1 word each clock cycle
    - DDR – double data rate – transfers 2 words each clock cycle
    - DDR2, DDR3, DDR4 – transfer 4,8,16 words each array access
  - Medium speed
  - Highest density
  - Used as main memory



Technology
Memory

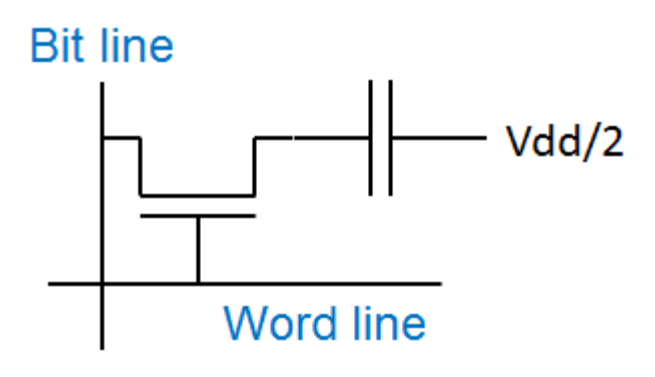
- SDRAM – Synchronous Dynamic Random Access Memory
  - Write
    - All Word lines low
    - Read Enable (RE) disabled (low)
    - Place B0, B1, B2, B3 on inputs
    - Pull write enable bar (WE) low
    - Strobe the desired word line high
    - Bit lines write to the bit cell capacitors

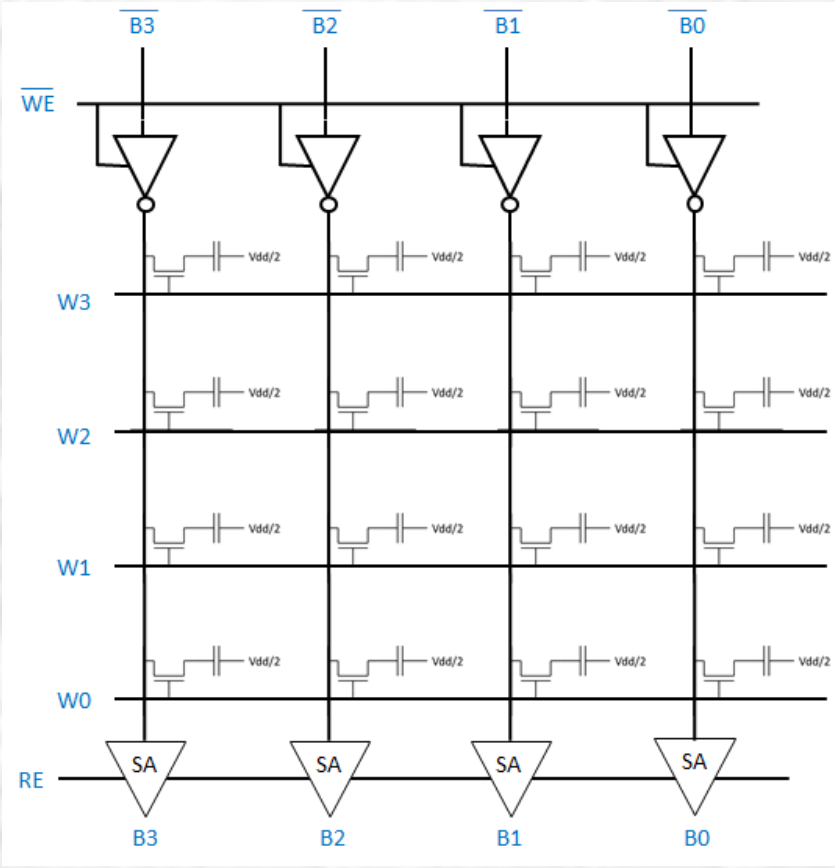




Technology
Memory

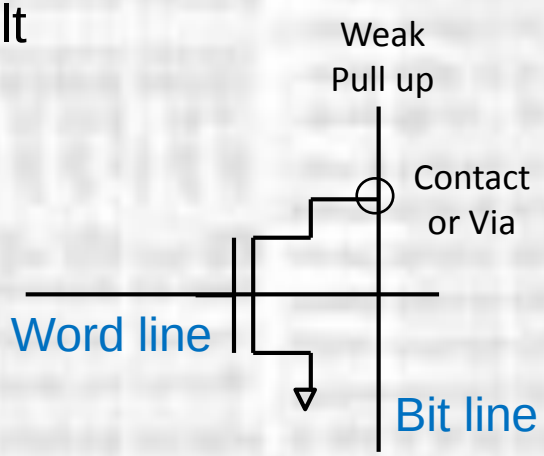
- SDRAM – Synchronous Dynamic Random Access Memory
  - Read
    - All Word lines low
    - Write enable bar (WE) high
      - inverters tristated
    - Read Enable (RE) high
    - Strobe the desired word line high
    - Sense amplifiers read the value of the capacitors





Technology
Memory

- ROM – Read Only Memory
  - Memory cell (1 bit) is based on whether a MOSFET is or is not connected between the bit line and word line
  - The MOSFET structure is always part of the bit cell
    - It can be connected to the bit line through a contact or via (via ROM)
    - It can be disable by removing the S/D diffusion (diffusion ROM)
  - Cannot be modified once the part is built
  - Typically used as Boot memory or to hold chip configuration data

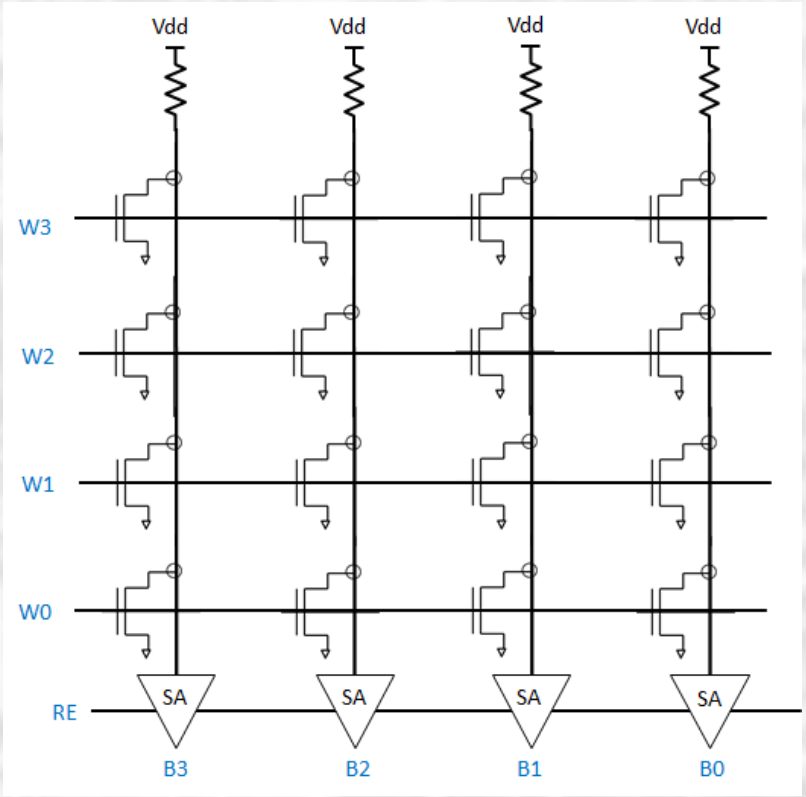


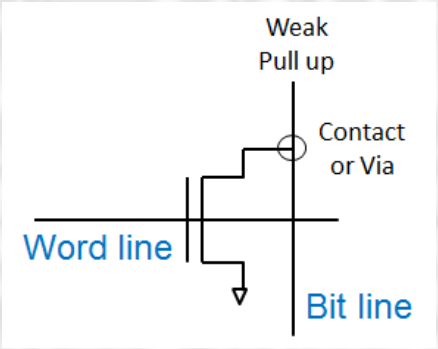
Technology
Memory

- ROM – Read Only Memory
  - Read
    - All Word lines low
    - Read Enable (RE) high
    - Strobe the desired word line high
    - Sense amplifiers read the value of the bit lines

    If connected – will read a “0”

    If not connected – will read a “1”



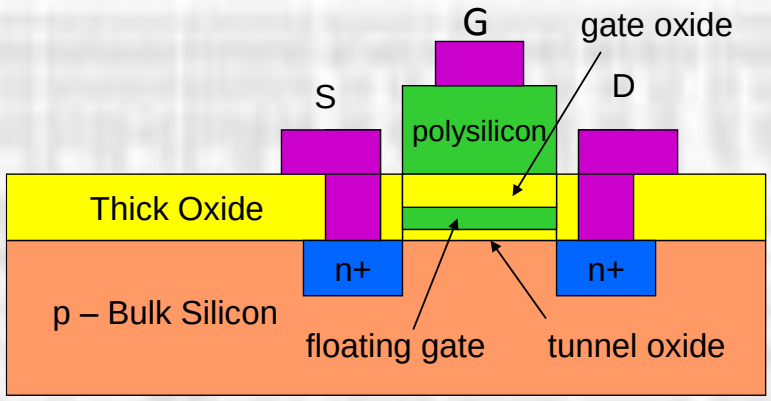


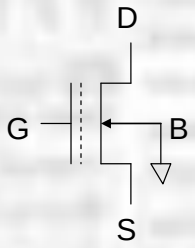
Technology
Memory

- Flash Memory
  - Memory cell (1 bit) is based on charge stored on a floating capacitor
    - The capacitor modifies the threshold voltage of a MOSFET
      - with negative charge stored – need higher gate voltage to turn on the MOSFET
    - Creates 2 possible threshold voltages

      $V_{th\ High}$ is required to turn on the MOSFET if charge is stored

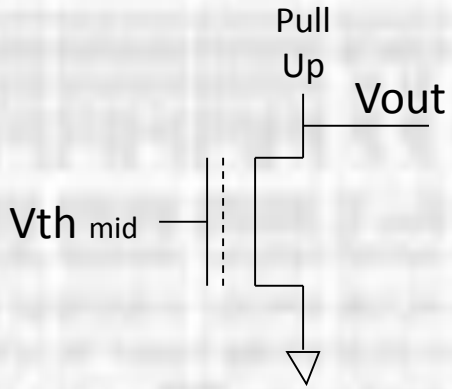
      $V_{th\ Low}$ is required to turn on the MOSFET if no charge is stored





Technology
Memory

- Flash Memory
  - Cell write
    - High voltage process that allows electrons to be injected into (erase) or tunnel out of (write) the floating gate
    - Once the high voltage is removed the electrons are trapped
  - Cell read
    - Place a voltage on the gate midway between Vth_{High} and Vth_{Low}
    - Use the circuit to determine if the MOSFET is on or off



If charge stored on capacitor (erased)

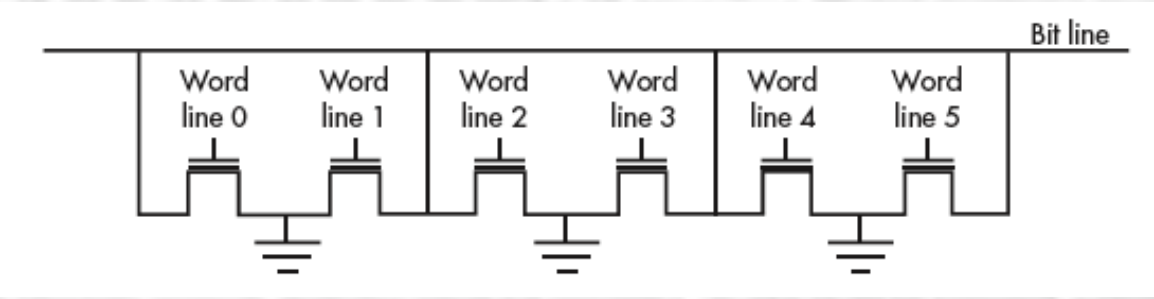
$Vth_{mid} < (Vth = Vth_{High}) \rightarrow$ Vout = high $\rightarrow$ "0"

If no charge stored on capacitor (programmed)

$Vth_{mid} > (Vth = Vth_{Low}) \rightarrow$ Vout = low $\rightarrow$ "1"

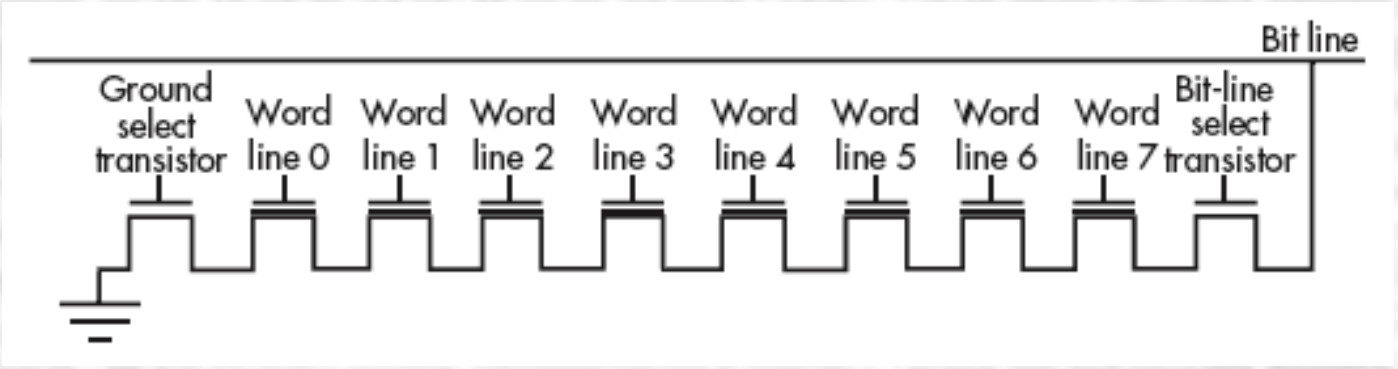
Technology
Memory

- Flash Memory
  - NOR Flash
    - Block Erase
    - Less dense than Nand Flash (extra wires and connections)
    - Slower sequential access than Nand Flash
    - Fast (true) random access
    - Used as program memory



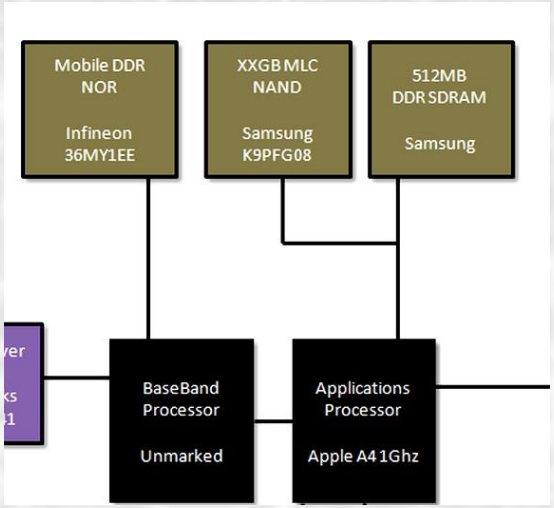
Technology
Memory

- Flash Memory
  - NAND Flash
    - Block Erase
    - More dense than Nor Flash
    - Faster sequential access than Nor Flash
    - Random access requires additional parts and time
    - Used as file storage memory (Flash Drives)



Technology
Memory

- Flash Memory
  - Shadowing
    - Store large amounts of program and data in Nand Flash
    - At boot, copy a portion of the Nand memory into SRAM or SDRAM
    - Use the SRAM as the processor program and data memory
    - As additional program or data are needed – swap out a portion of the SRAM/SDRAM



Technology
Memory

- Flash memory
  - XIP – Execute in Place
    - Execute directly out of NOR flash
      - Nor Flash densities are growing rapidly
      - Nor Flash speeds are fast enough to support the memory hierarchy
      - Requires a caching system

Technology
Memory

- Access
  - Many variations, but all basically the same
  - Two Buses
    - Address Bus
      - Address width (# of bits) is $\log_2 m$ where m is the number of memory locations
      - Uni-directional – only written by processor, read by memory
    - Data Bus
      - Data width (# of bits) is set by the memory/system interface
      - Bi-directional – both memory and processor can R/W this bus
  - 3 Control Signals
    - Chip Enable
      - Used to choose one of several possible memory devices
    - Read
      - Asserted by the procesor to signal a read operation
    - Write
      - Asserted by the processor to signal a write operation

Technology

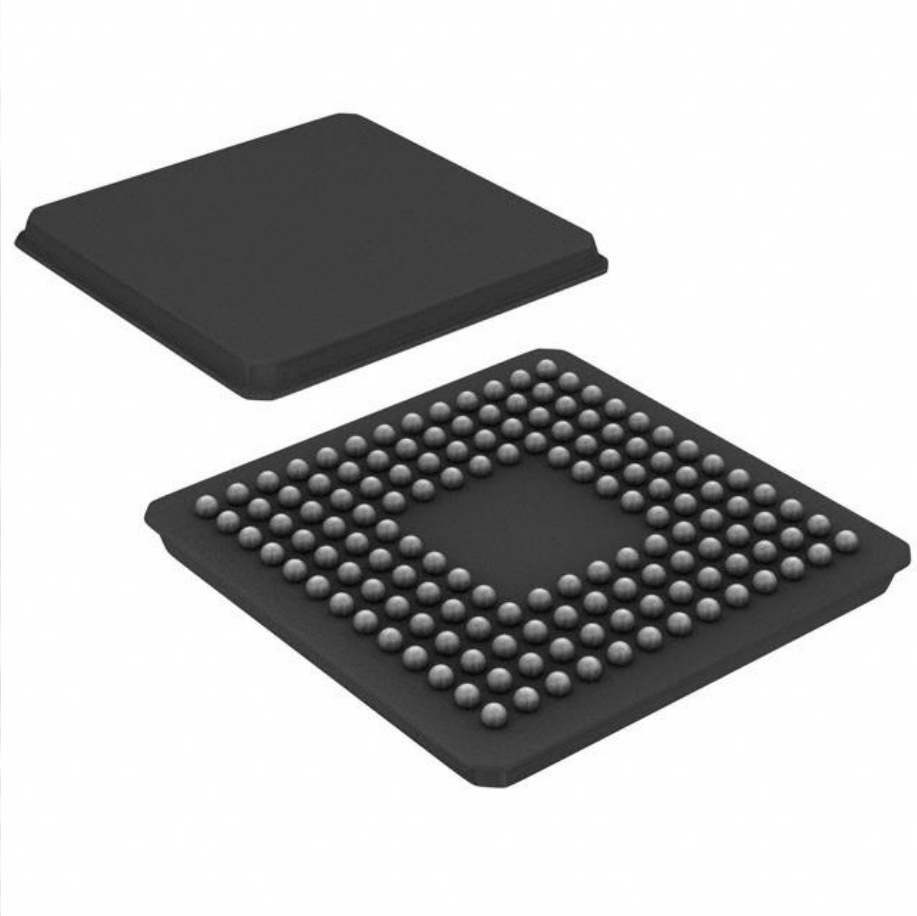
Memory

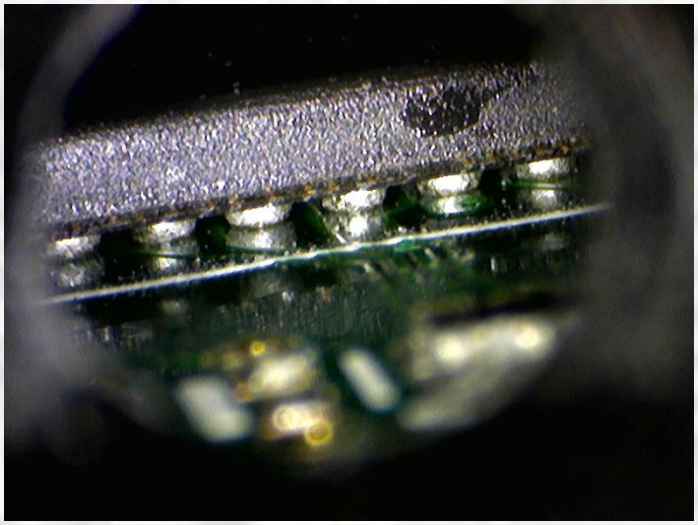
- Access
  - Simplified Process
  - Read
    - Processor selects the appropriate CE
    - Processor places the desired read address on the address bus
    - Processor asserts the read signal
    - Memory fetches the data and places it on the data bus
    - Processor reads the data bus
  - Write
    - Processor selects the appropriate CE
    - Processor places the desired write address on the address bus
    - Processor places the data on the data bus
    - Processor asserts the write signal
    - Memory reads the data on the data bus and stores it

Technology

Packaging

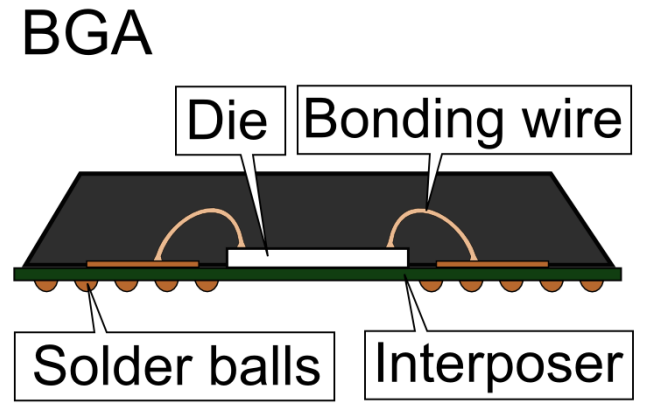
- Ball Grid Array

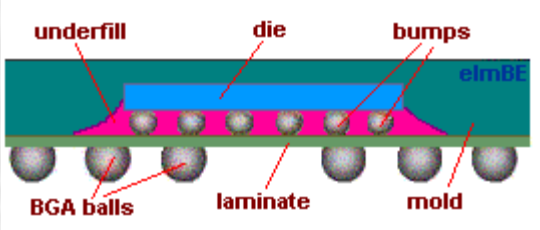




Technology
Packaging

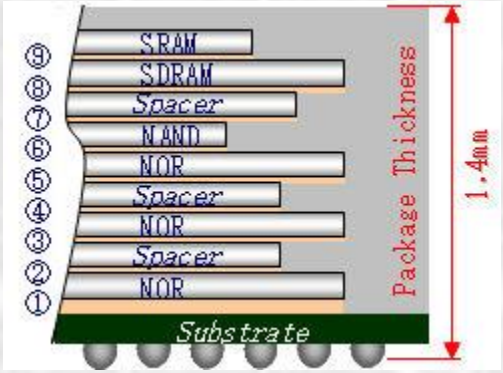
- Ball Grid Array
  - Bonded vs. Flip Chip

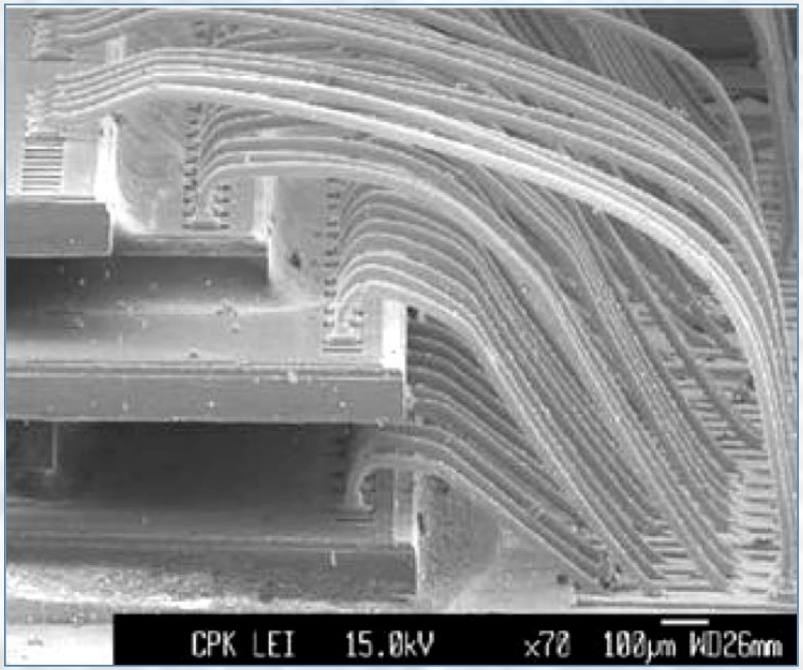




Technology
Packaging

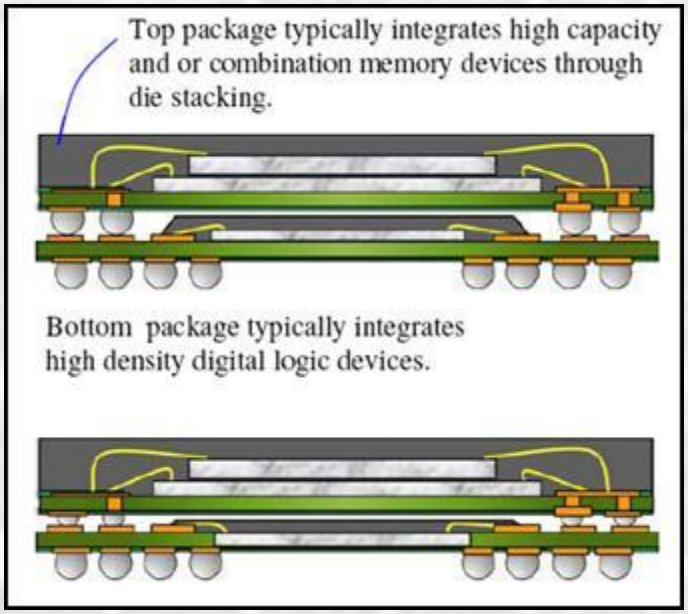
- Stacked Package





Technology
Packaging

- PoP Package



Technology

Packaging

- Silicon Through Via Package