ECE520 – VLSI Design

Lecture 21: Memories

Payman Zarkesh-Ha

Office: ECE Bldg. 230B
Office hours: Wednesday 2:00-3:00PM or by appointment
E-mail: pzarkesh@unm.edu

Review of Last Lecture

- **Power Distribution Network**
  - **IR drop**
  - **Switching noise**
  - **Decoupling capacitor**
  - **Electromigrartion**
- **I/O Circuits**
  - **Output buffer**
  - **Level shifters**
  - **Schmitt trigger**
  - **Tri-state outputs**
  - **ESD protection**

Today's Lecture

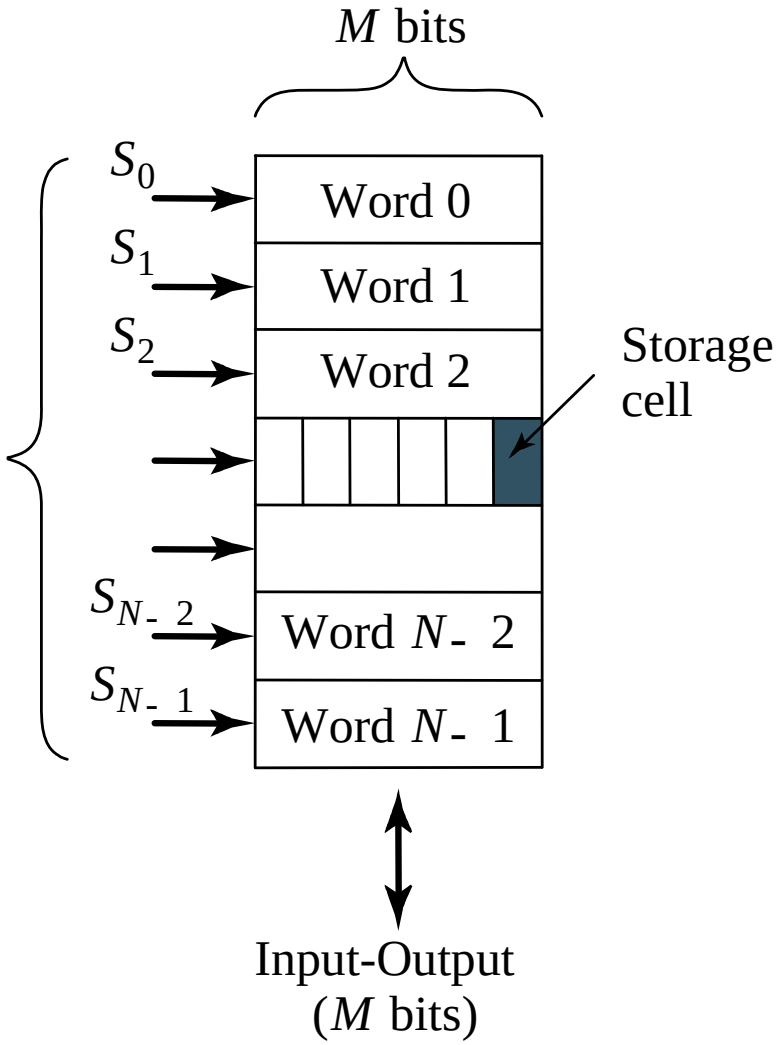
- **Memory architectures**
- **ROM**
  - **NOR ROMs**
  - **NAND ROMs**

Types of VLSI Memories

- **Register files**
  - Fastest and largest--sub 200ps access times possible
  - Readily implemented with multiple ports
- **SRAM**
  - Next fastest--sub 1ns access times possible
  - This includes cache memory--they are all SRAM based
  - Volatile, but retains state as long as power is applied
- **DRAM**
  - Quite slow > 50ns access times
  - Volatile--state retained for ms and must be refreshed often
- **ROM**
  - Mask programmable Read Only Memory
- **Non-volatile memories**
  - Slow--50ns access times
  - EPROM, EEPROM, Flash
  - Emerging technologies: Ferroelectric, ferromagnetic…

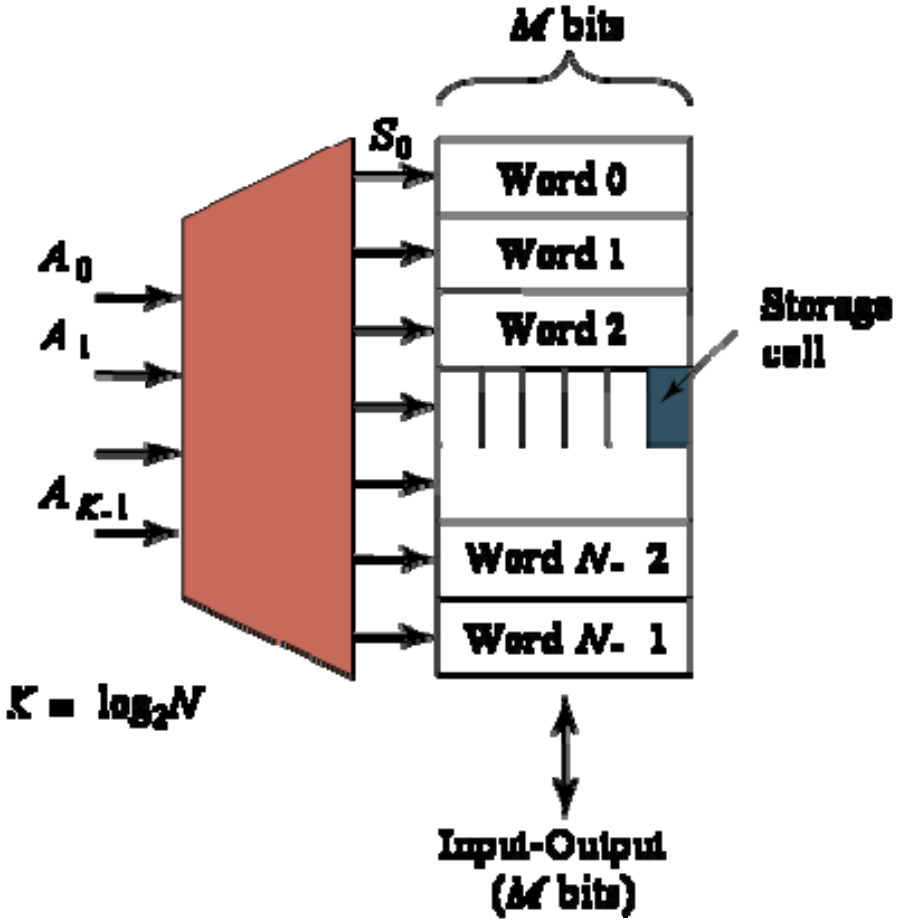
A Generic Memory Architecture

- Words are horizontal and accessed by applying the appropriate word line
- Each cell in a row is a bit in the corresponding word
- This is not practical. Why?



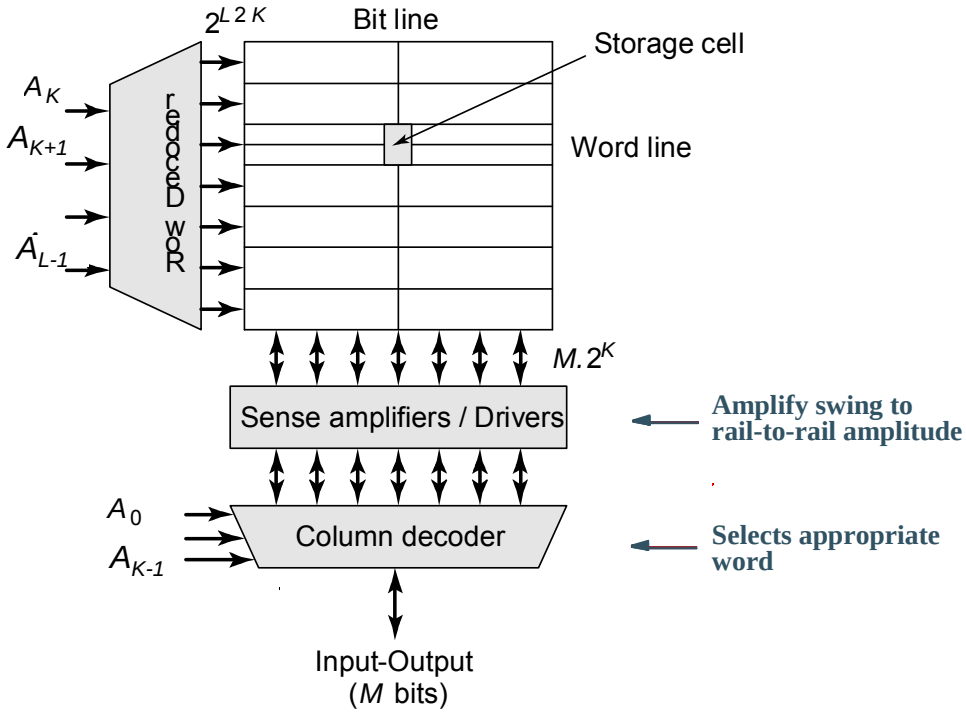
Memory with Decoder

- Decoder reduces the number of select signal significantly
- In this case $k=\log_2 N$
- Is this a reasonable architecture for 256 8-bit words?



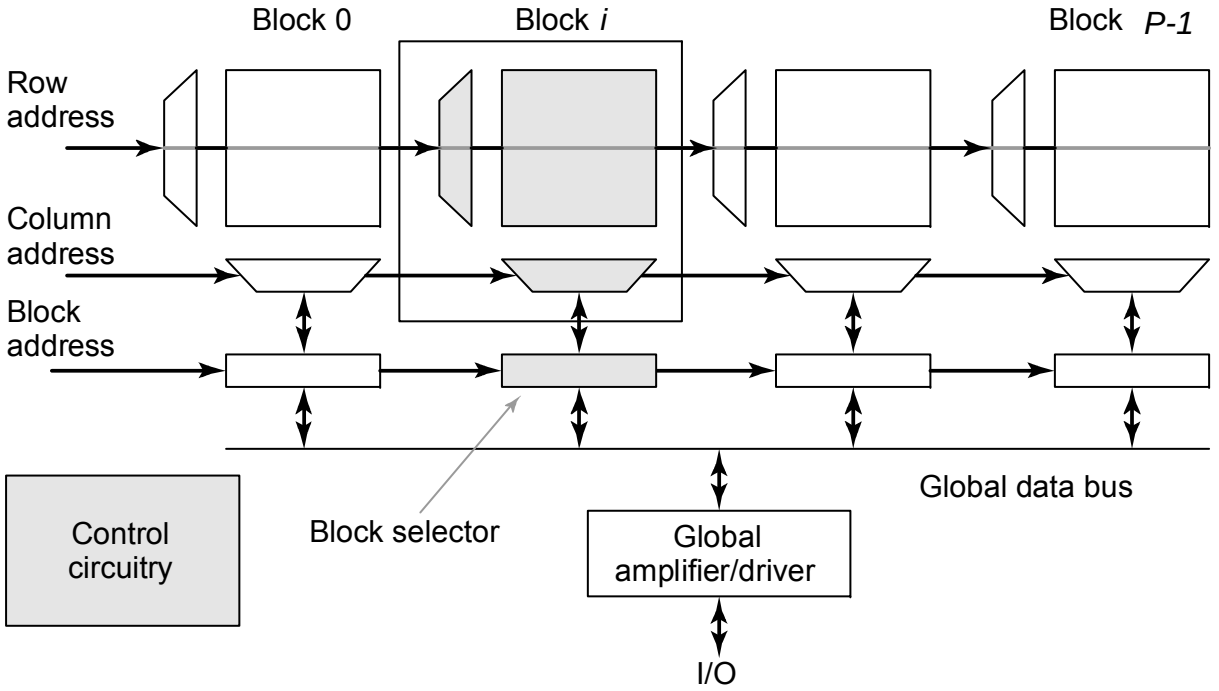
Real Memory Block Organization

- To improve Aspect Ratio, memory arrays are organized so that vertical and horizontal dimensions are the same order
- The row address enables one row of the memory for R/W, while the column address picks one particular word from the address

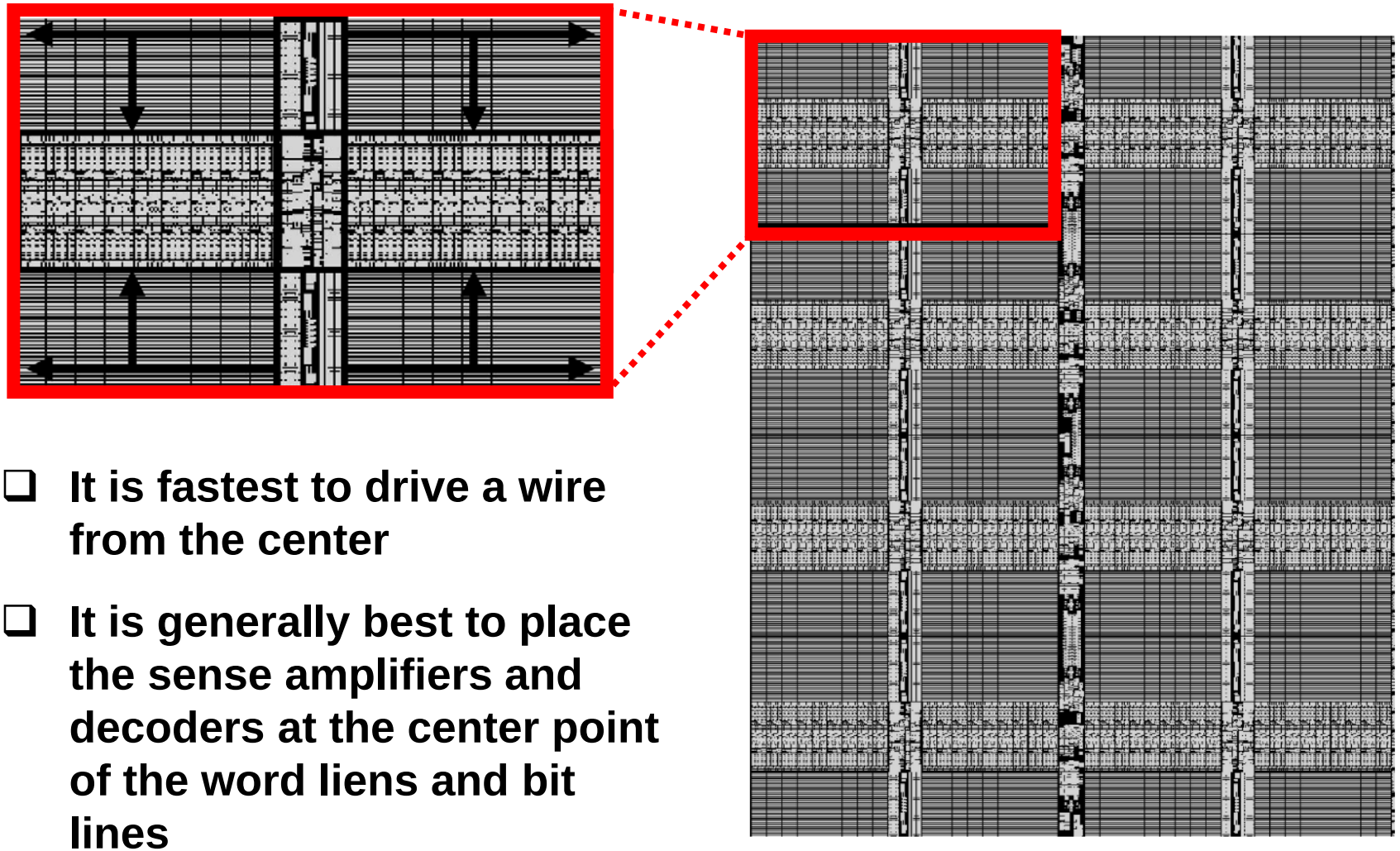


Memory Banking (Hierarchical Architecture)

- The length of the local word and bit lines is kept within bounds, resulting faster access time
- Only the addressed block is activated resulting in a substantial power saving. Power dissipation is a major concern in very large memories.



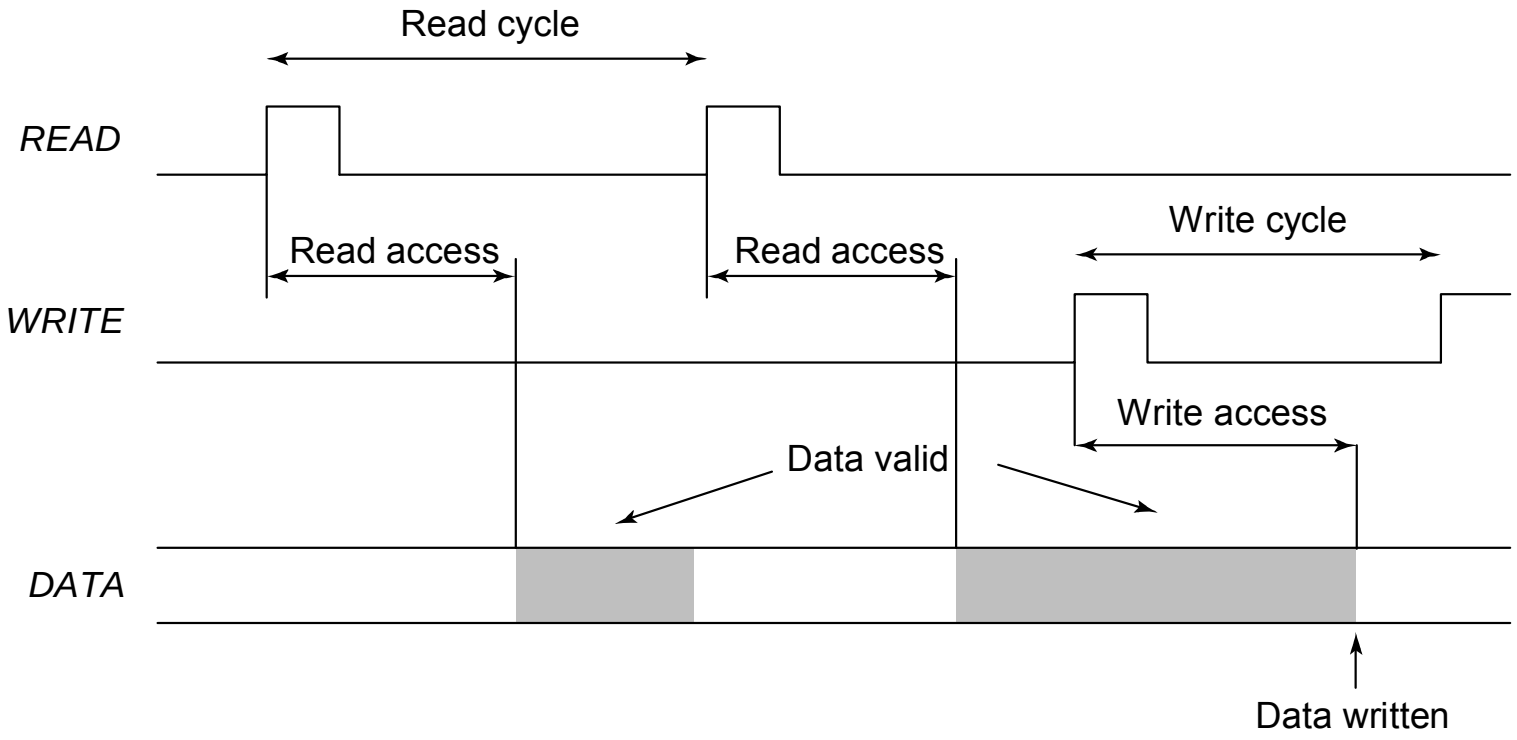
Example: Physical Memory Organization



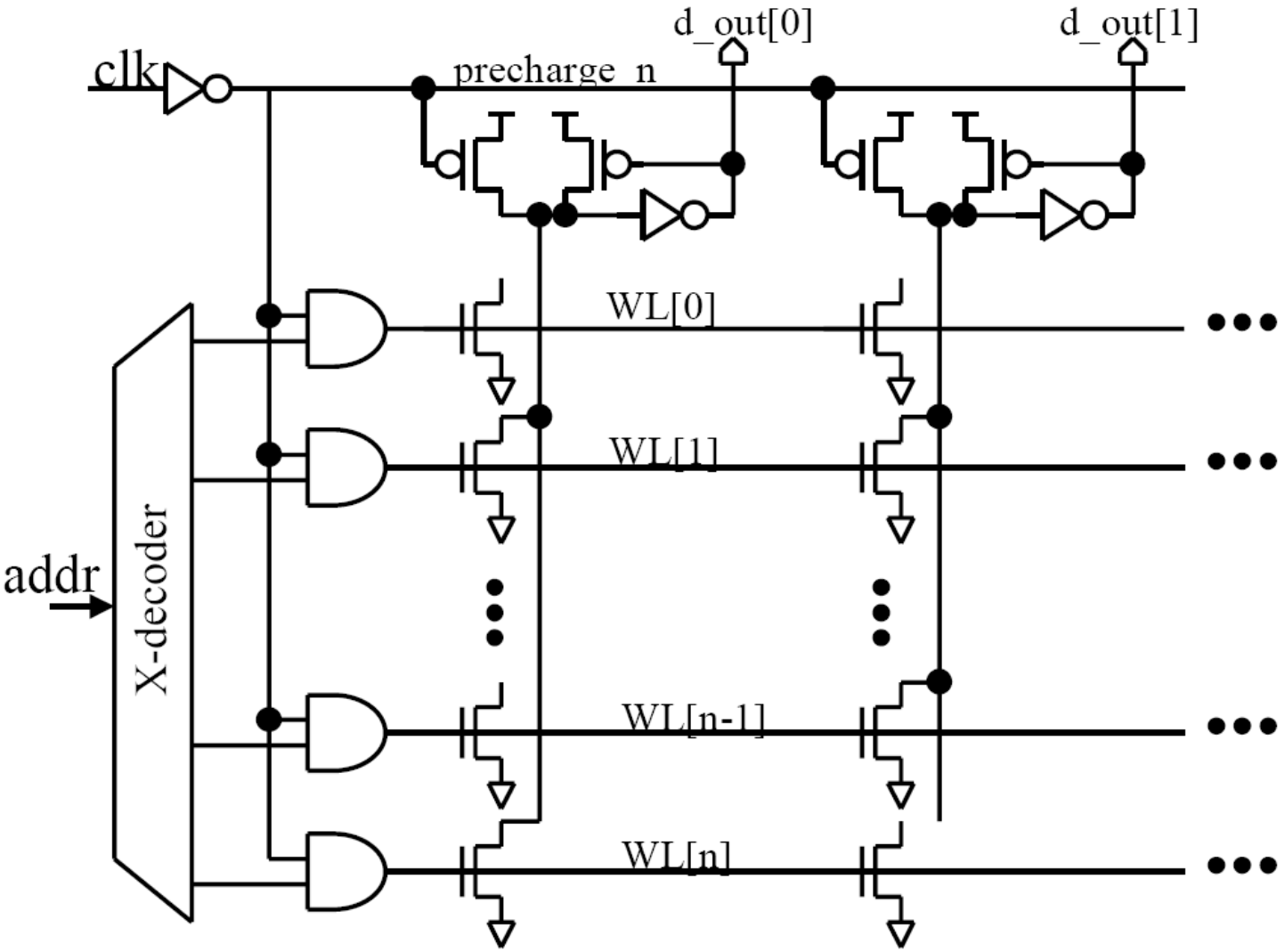
- It is fastest to drive a wire from the center
- It is generally best to place the sense amplifiers and decoders at the center point of the word liens and bit lines

Memory Read and Write Timing

- **Memory timing parameters includes: *read-access time*, *write-access time*, *read-cycle time*, and *write-cycle time***



Read Only Memory (ROM)

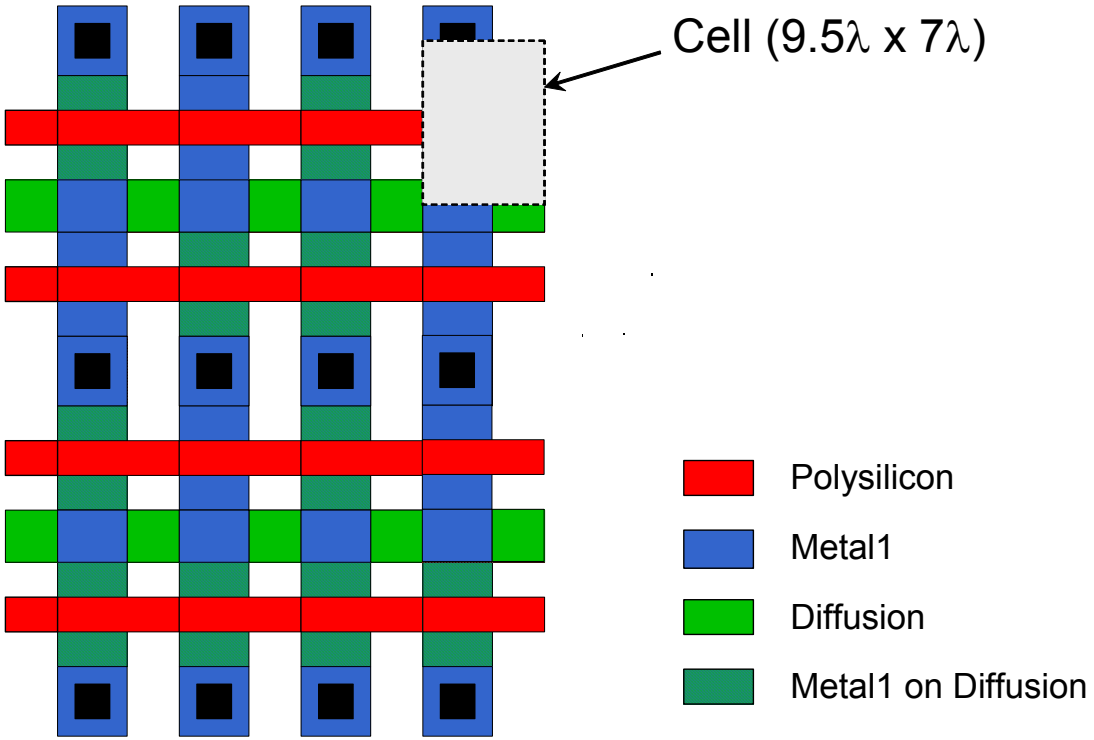


Read Only Memory (ROM)

- **ROM operation**
  - Bit lines are pre-charged during clock high in this example
  - Word lines are enabled during the low phase of the clock
  - Note that the cells are simple domino circuits - just a simple pull-down
  - PMOS keepers keep the bit lines from discharging due to leakage if the bit read is a logic one, i.e., the no pull down case
- **ROM programming**
  - Logic 0 or 1 in a cell is programmed during manufacture by including a drain connection or not
  - No drain connection, the cell cannot pull low
  - Drain connection exists, when selected the cell will pull low
  - Programming Schemes: 1) Diffusion programming (allows contact sharing), 2) Contact/via programming--generally not as small and 3) Vt implant programming

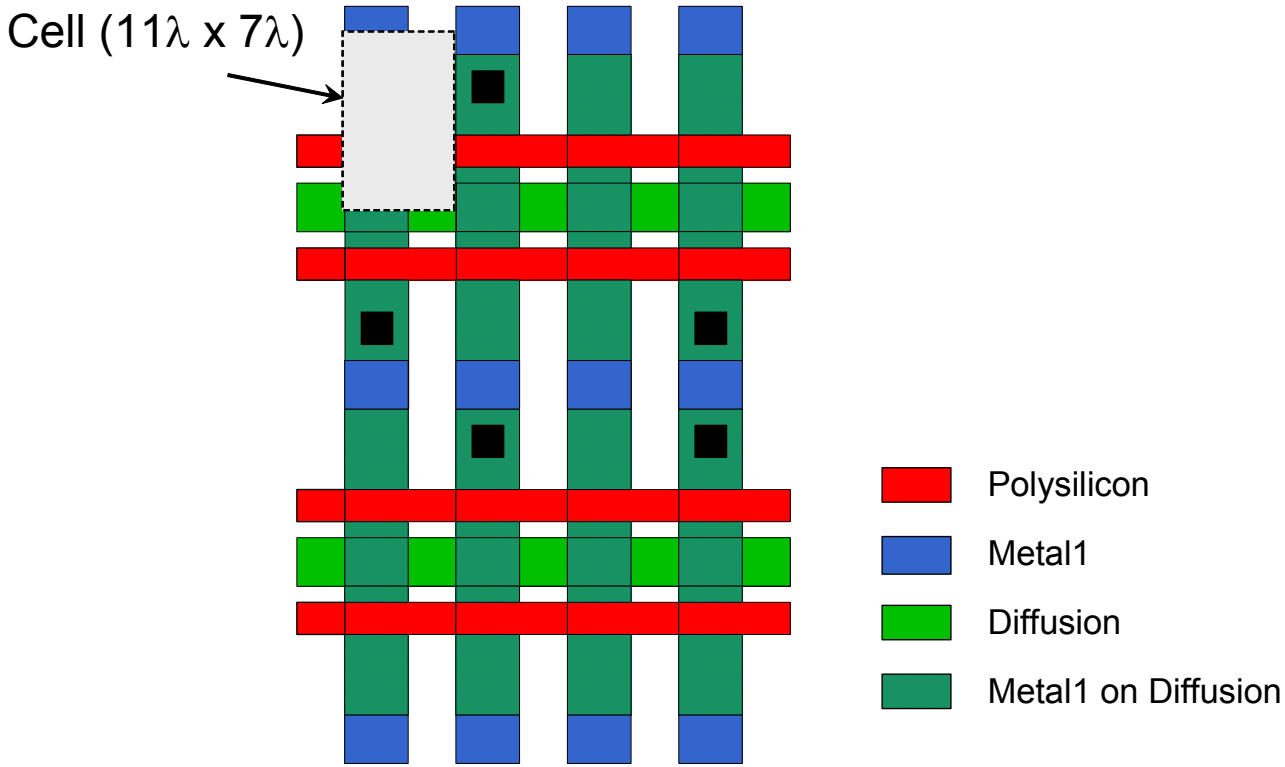
ROM Programming Using Active Layer

- M1 is used for the bit lines
- Poly is used for the word lines
- Diffusion is either absent or present to form for pull-down or not
- Contacts shared between adjacent cells



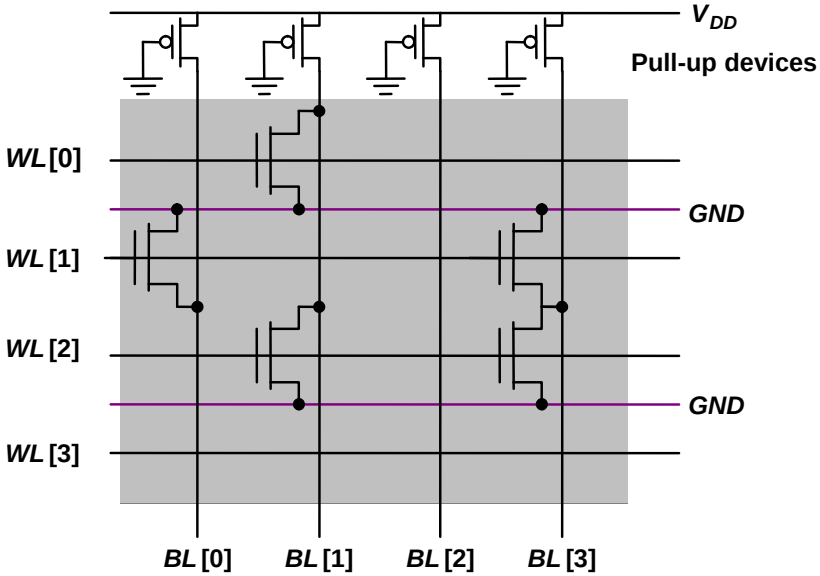
ROM Programming Using Contacts

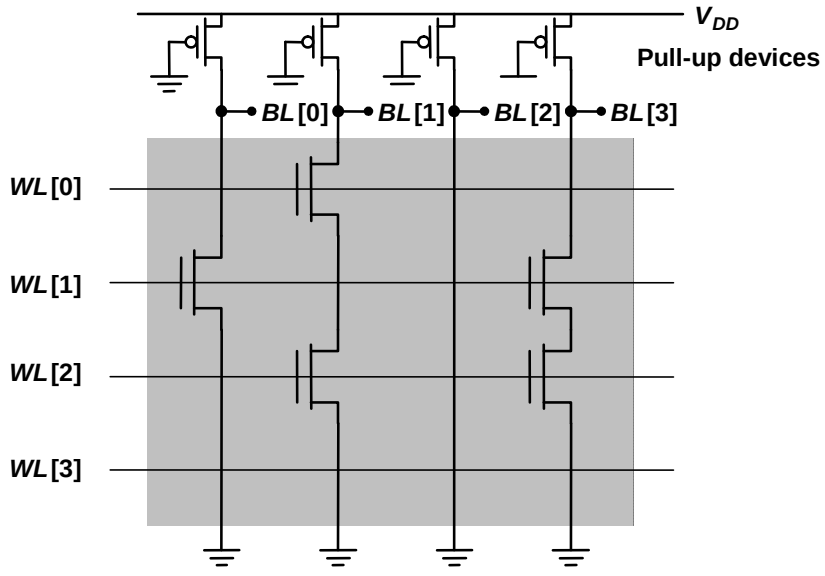
- M1 is used for the bit lines
- Poly is used for the word lines
- Contact is either absent or present to form for pull-down or not
- Contacts cannot be shared – results in larger layout area



NOR versus NAND ROM

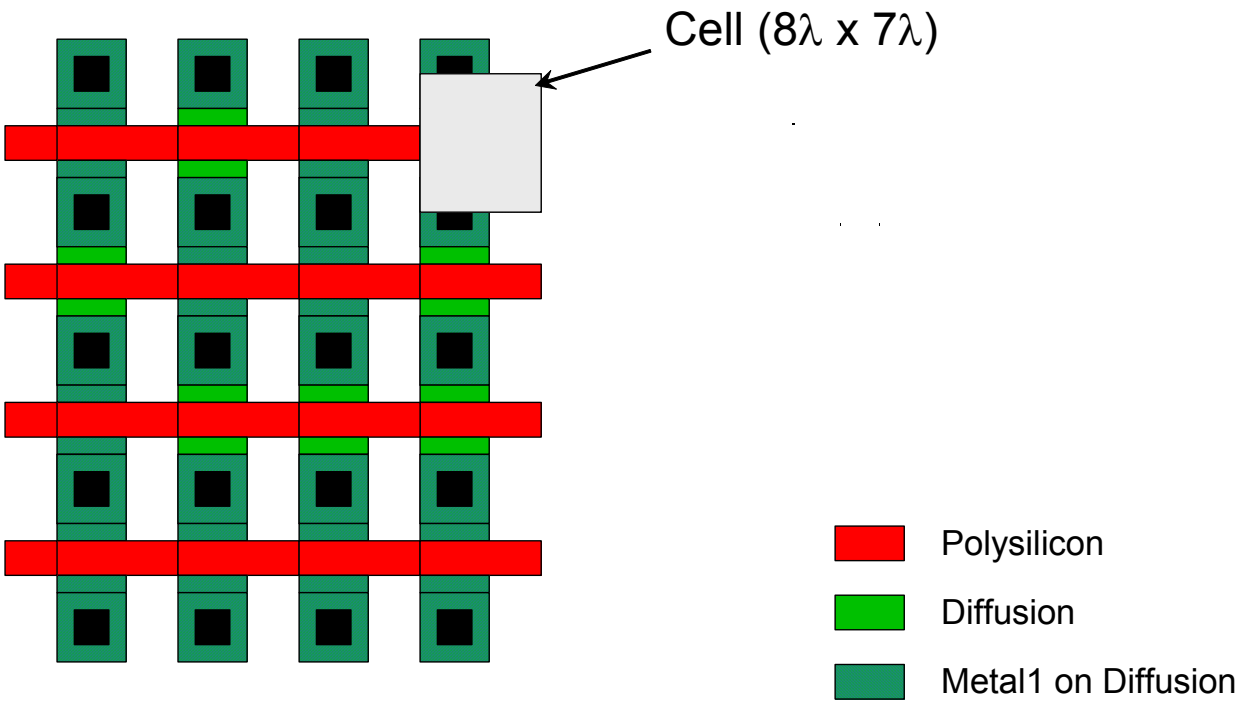
- NAND ROM doesn't need any connection to supply voltage, which results in smaller area
- NAND ROM operates in reverse-logic mode, i.e. all word lines are high by default except the selected row which is 0
- All transistors in the pull down chain must be on to produce 0
- If a transistor exists on the selected row it creates 1 (pull down will be off), otherwise it creases 0 (pull down is on)





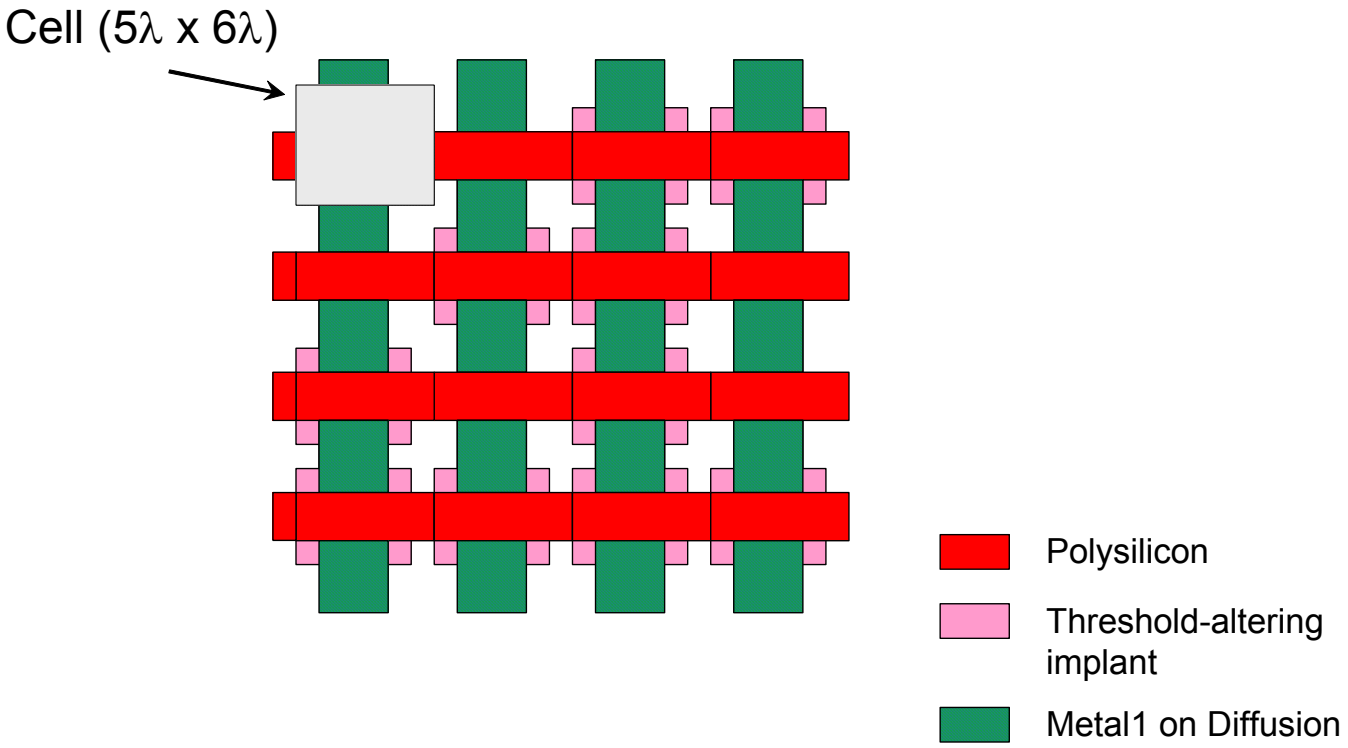
NAND ROM Programming Using Metal 1

- M1 is used for the bit lines and to selectively short circuit the transistors
- Poly is used for the word lines



NAND ROM Programming Using Implant

- Implant (n-type) can be used to selectively turns the devices into depletion mode transistor, which is always on, regardless of applied word line voltage
- This has even less layout area



Problems with NAND ROMs

- **The improved density of NAND ROMs comes at a price:**
  - **It is only useful for small memory arrays. why?**
  - **It results in a considerable performance loss. Why?**