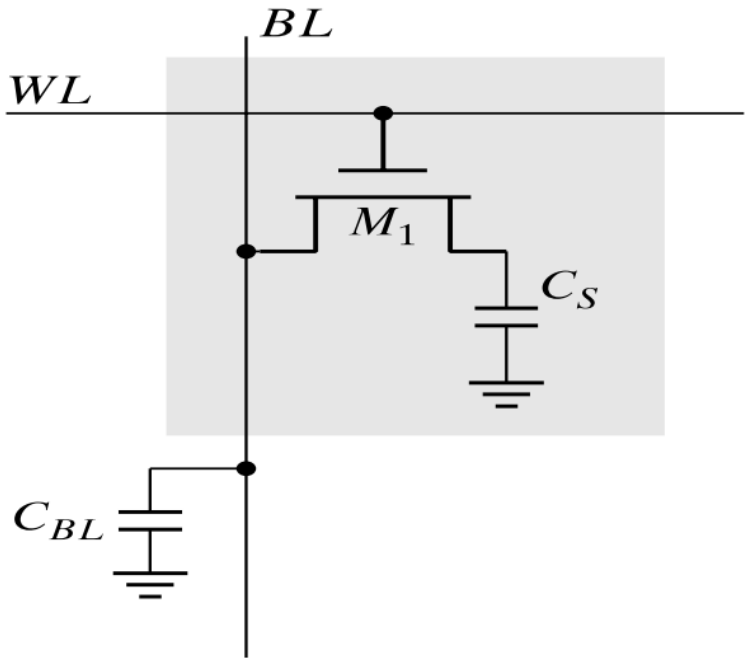
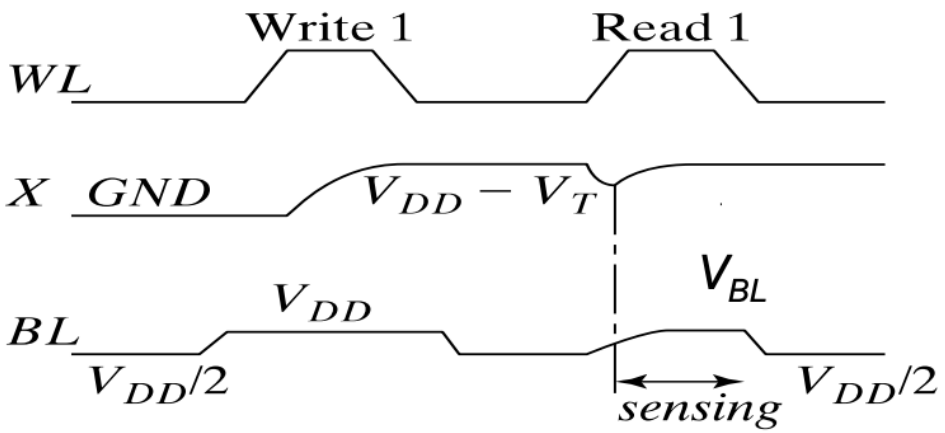
Discussion Section 9

Sean Huang
April 2, 2021

Berkeley
UNIVERSITY OF CALIFORNIA

DRAM

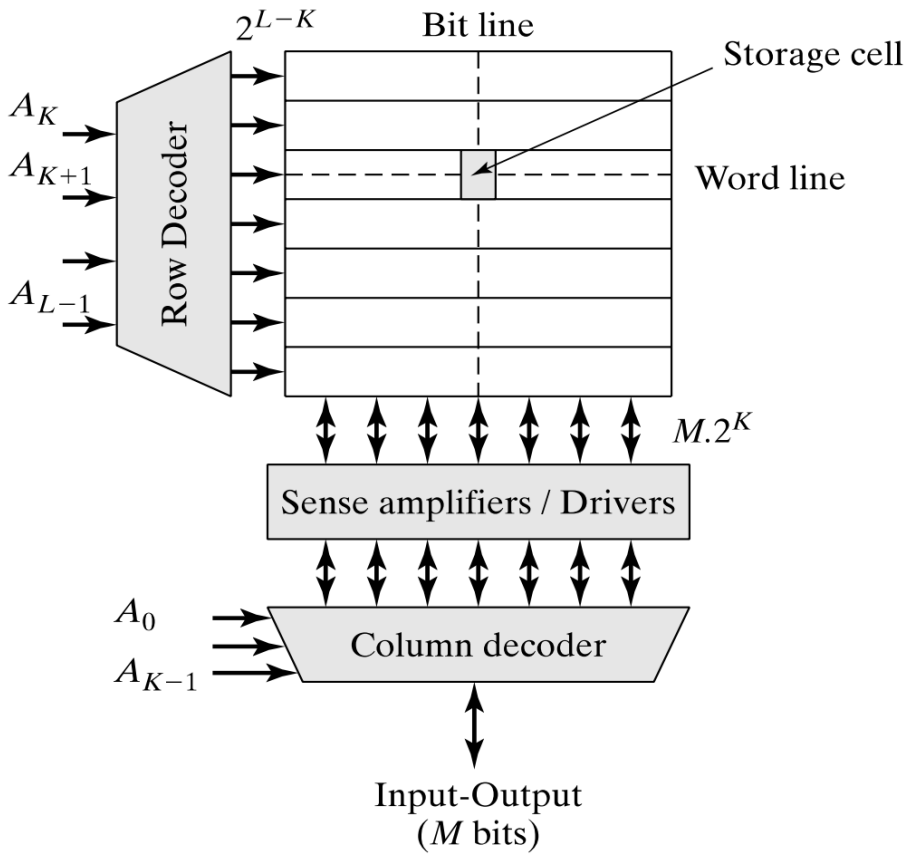




$V_{BIT} = 0$ or $(V_{DD} - V_T)$

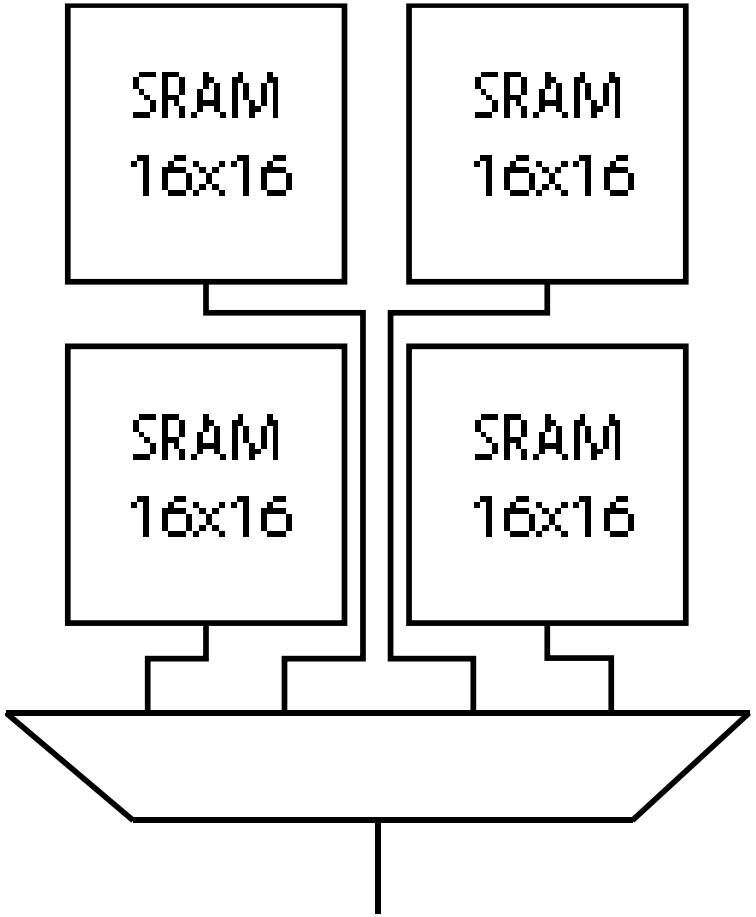
Memory Block

- **Word lines** used to select a row for reading or writing
- **Bit lines** carry data to/from periphery
- **Core** *aspect ratio* keep close to 1 to help balance delay on word line versus bit line
- **Address bits** are divided between the two decoders
- **Row decoder** used to select word line
- **Column decoder** used to select one or more columns for input/output of data



Large Memories

- Make larger SRAMs out of smaller SRAMs
- Each sub-SRAM has its own periphery circuits
- Need to map top-level address to each SRAM

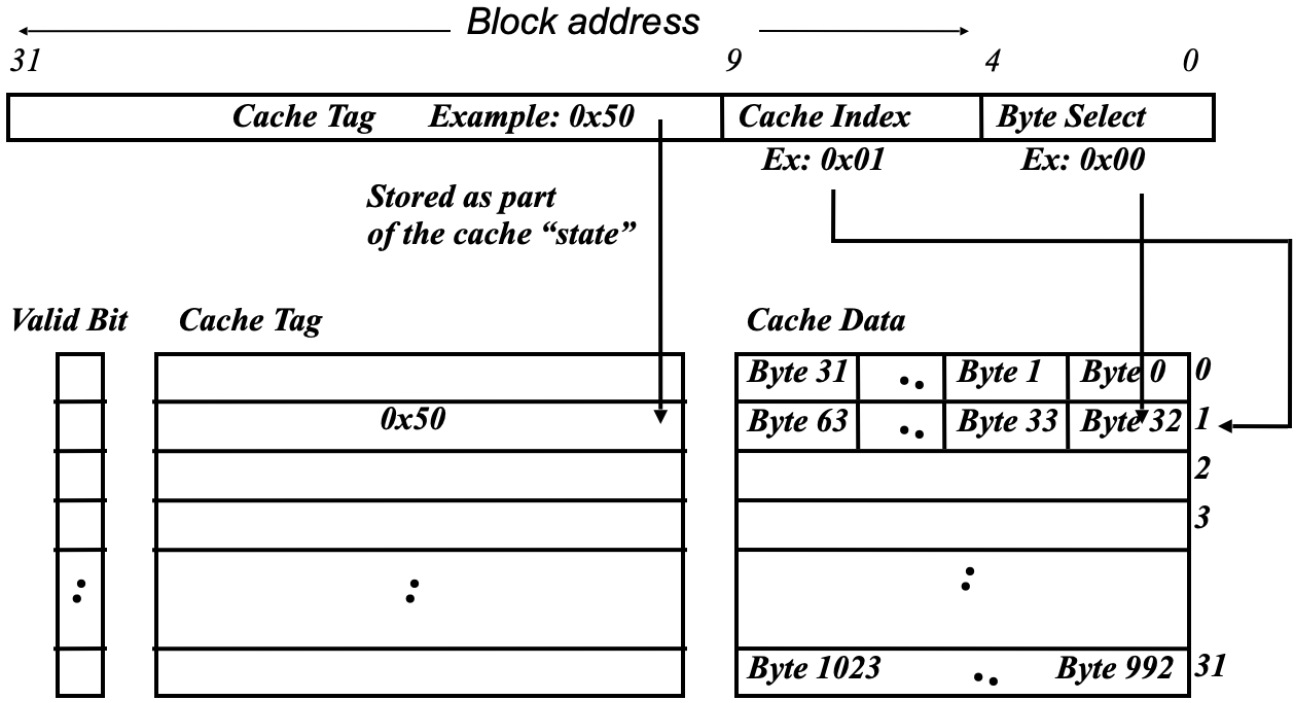


Cache Review

- Direct-mapped
- Fully-associative
- Set-assosciative

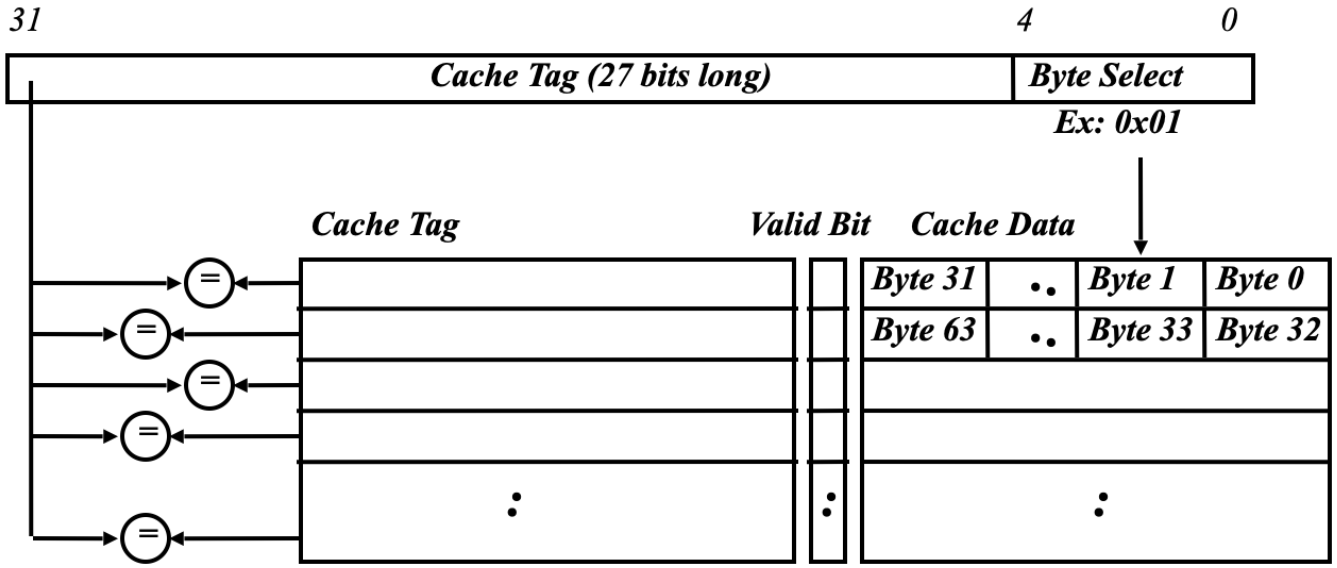
Direct-Mapped

- Each address maps to a position in cache block
- Address contains tag, cache controller checks against this for each read access
- On miss, cache fetches correct data from main memory



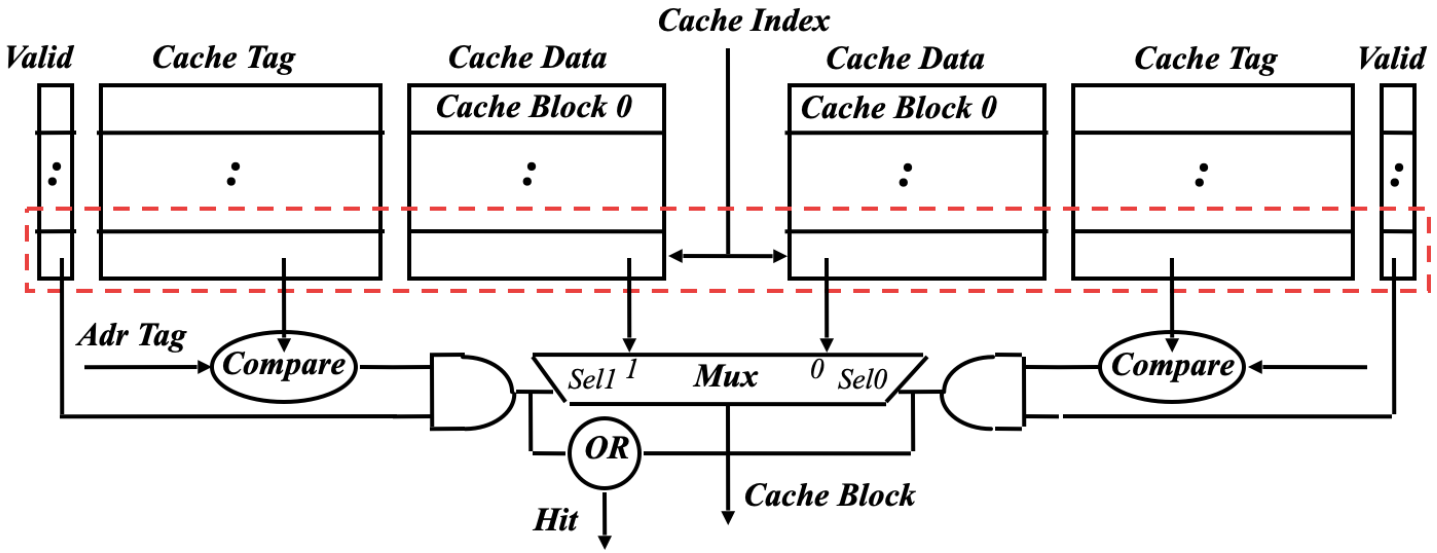
Fully Associative

- Any address can occupy any space in the cache
- Allows for more temporal locality if memory locations might not be mapped close together
- Tag is entire address, so large part of cache is tag



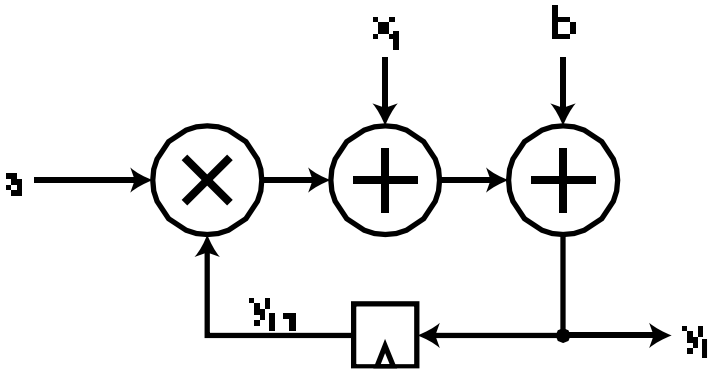
Set-Associative

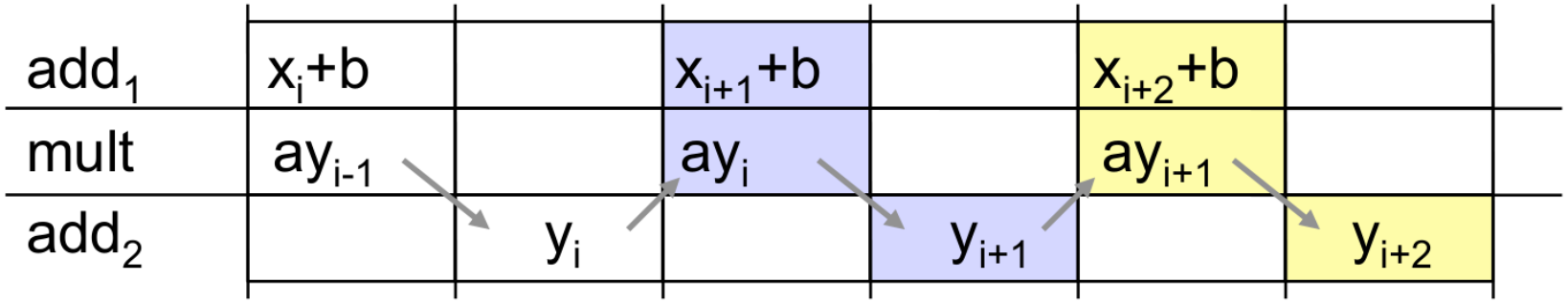
- In-between direct-mapped and fully associative
- Divide cache memory into "sets"
- Each memory location maps to a set, memory is fully associative within the set.
- Smaller tag than FA, may have better temporal locality than direct-mapped

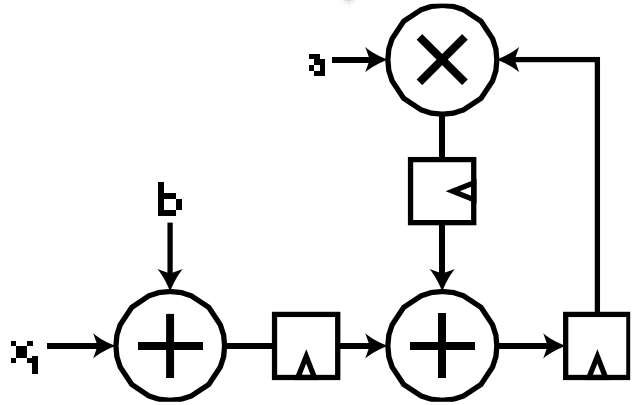


C-Slowing

- From the example in lecture, can reorder loop and introduce new delays to pipeline the computation

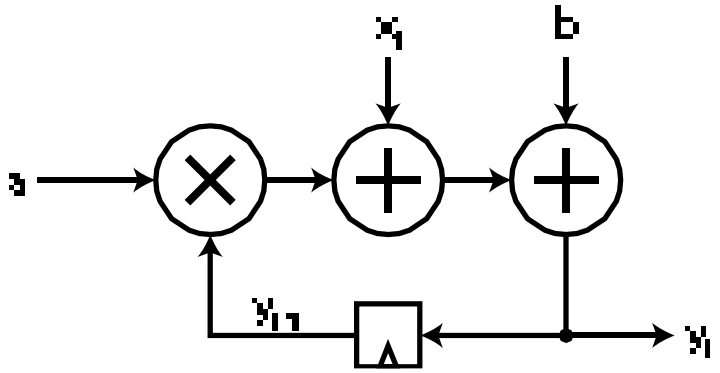


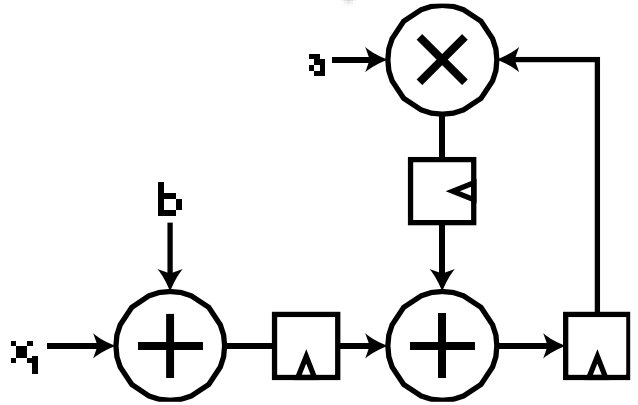




C-Slowing

- From the example in lecture, can reorder loop and introduce new delays to pipeline the computation
- Half of pipelined loop is idle, so we can queue another task here





add_1	x+b	x+b	x+b	x+b	x+b	x+b
mult	ay	ay	ay	ay	ay	ay
add_2	y	y	y	y	y	y

Loop Unrolling

- By replicating the loop a few times, we can take multiple inputs in parallel and generate multiple outputs in parallel
- Beware the long critical path from the chain of operations