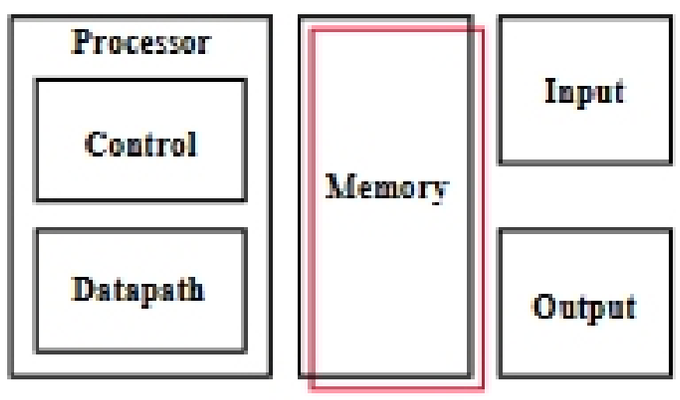
ECE4680
Computer Organization and Architecture

Memory Hierarchy

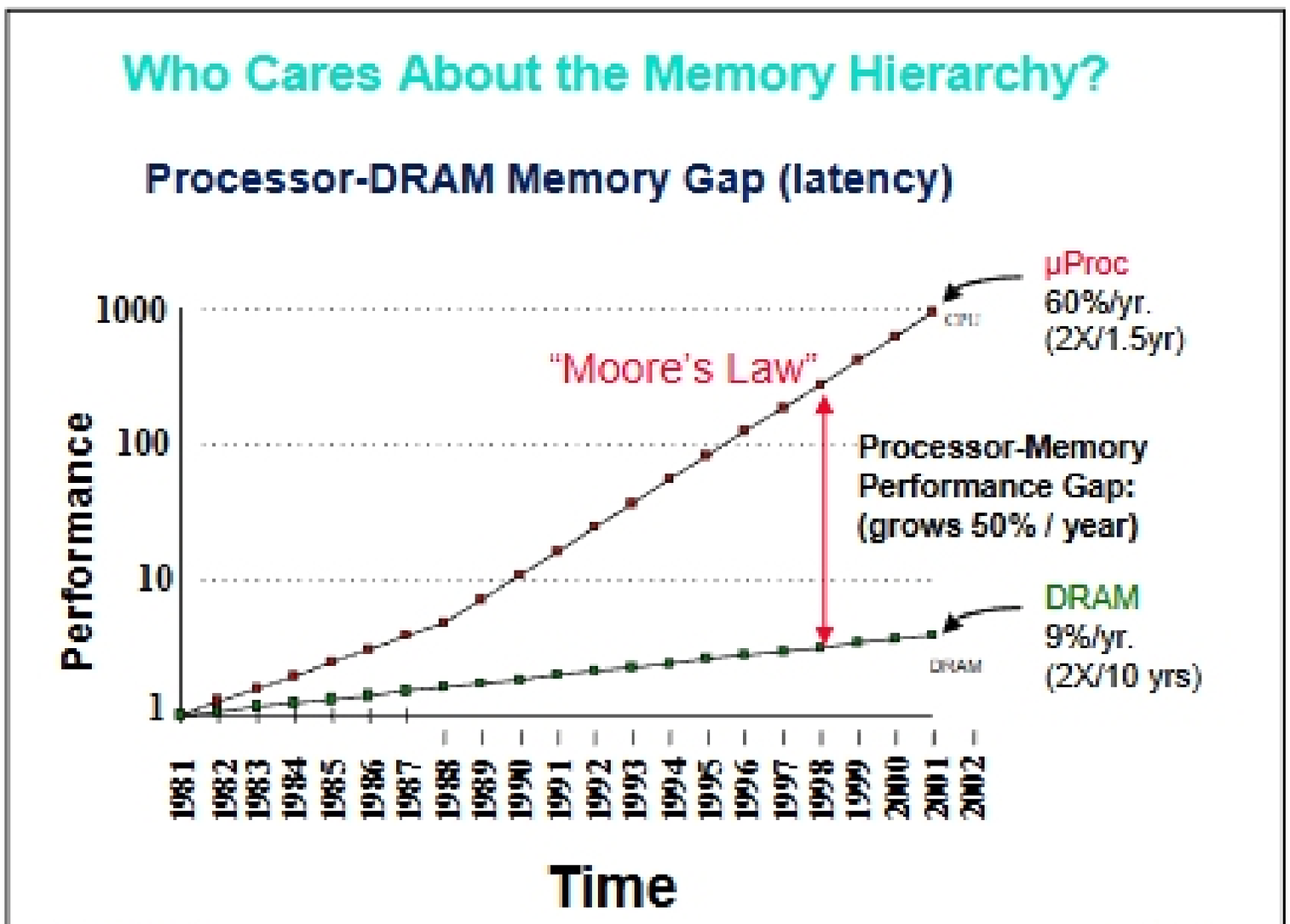
The Big Picture: Where are We Now?

- The Five Classic Components of a Computer

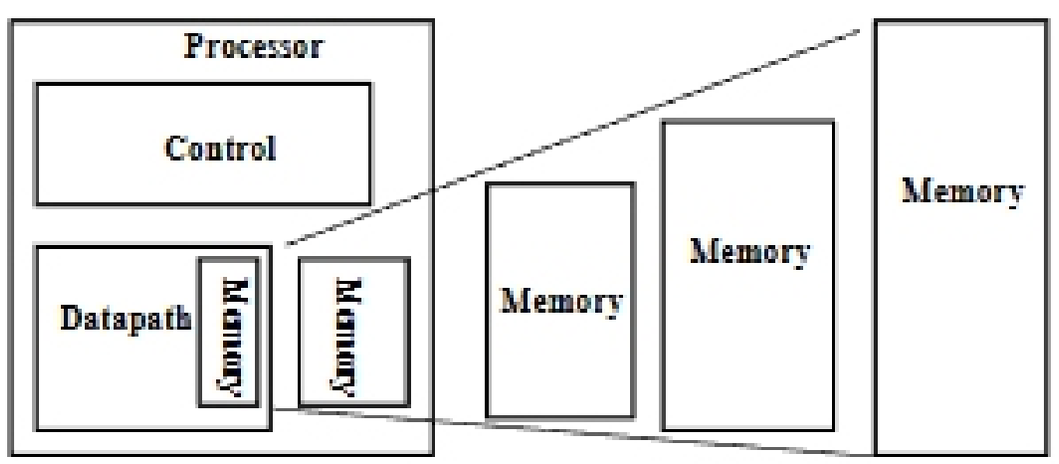


- Today's Topic: Memory System

Who Cares About the Memory Hierarchy?

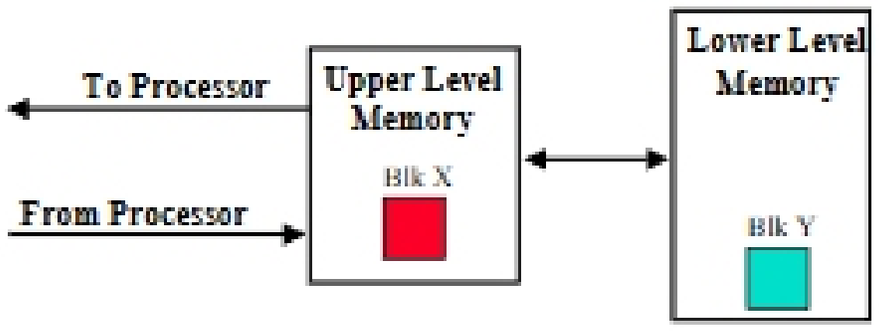


An Expanded View of the Memory System



Memory Hierarchy: Principles of Operation

- At any given time, data is copied between only 2 adjacent levels:
  - Upper Level: the one closer to the processor
    - Smaller, faster, and uses more expensive technology
  - Lower Level: the one further away from the processor
    - Bigger, slower, and uses less expensive technology
- Block:
  - The minimum unit of information that can either be present or not present in the two level hierarchy



Memory Hierarchy: Terminology

- Hit: data appears in some block in the upper level (example: Block X)
  - Hit Rate: the fraction of memory access found in the upper level
  - Hit Time: Time to access the upper level which consists of RAM access time + Time to determine hit/miss
- Miss: data needs to be retrieve from a block in the lower level (Block Y)
  - Miss Rate = 1 - (Hit Rate)
  - Miss Penalty: Time to replace a block in the upper level + Time to deliver the block to the processor
- Hit Time << Miss Penalty