Reconciling Repeatable Timing with Pipelining and Memory Hierarchy

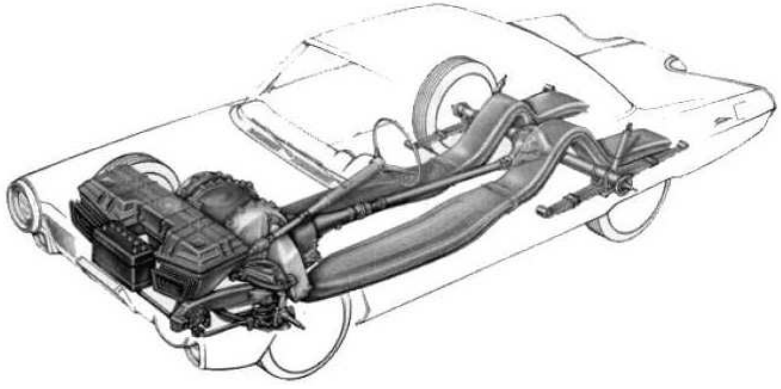
Stephen A. Edwards, Sungjun Kim (Columbia),
Edward A. Lee (UC Berkeley),
Hiren D. Patel (~~UC Berkeley~~ Waterloo),
and Martin Schoeberl (TU Vienna)

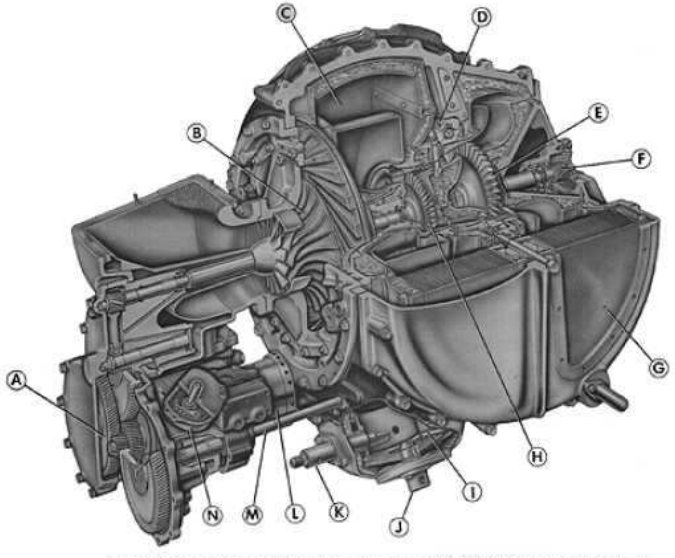
October 15, 2009

Timing Matters

1963 Chrysler Turbine Ghia

Ran on any fuel: gas, diesel, scotch, Chanel #5, tequila



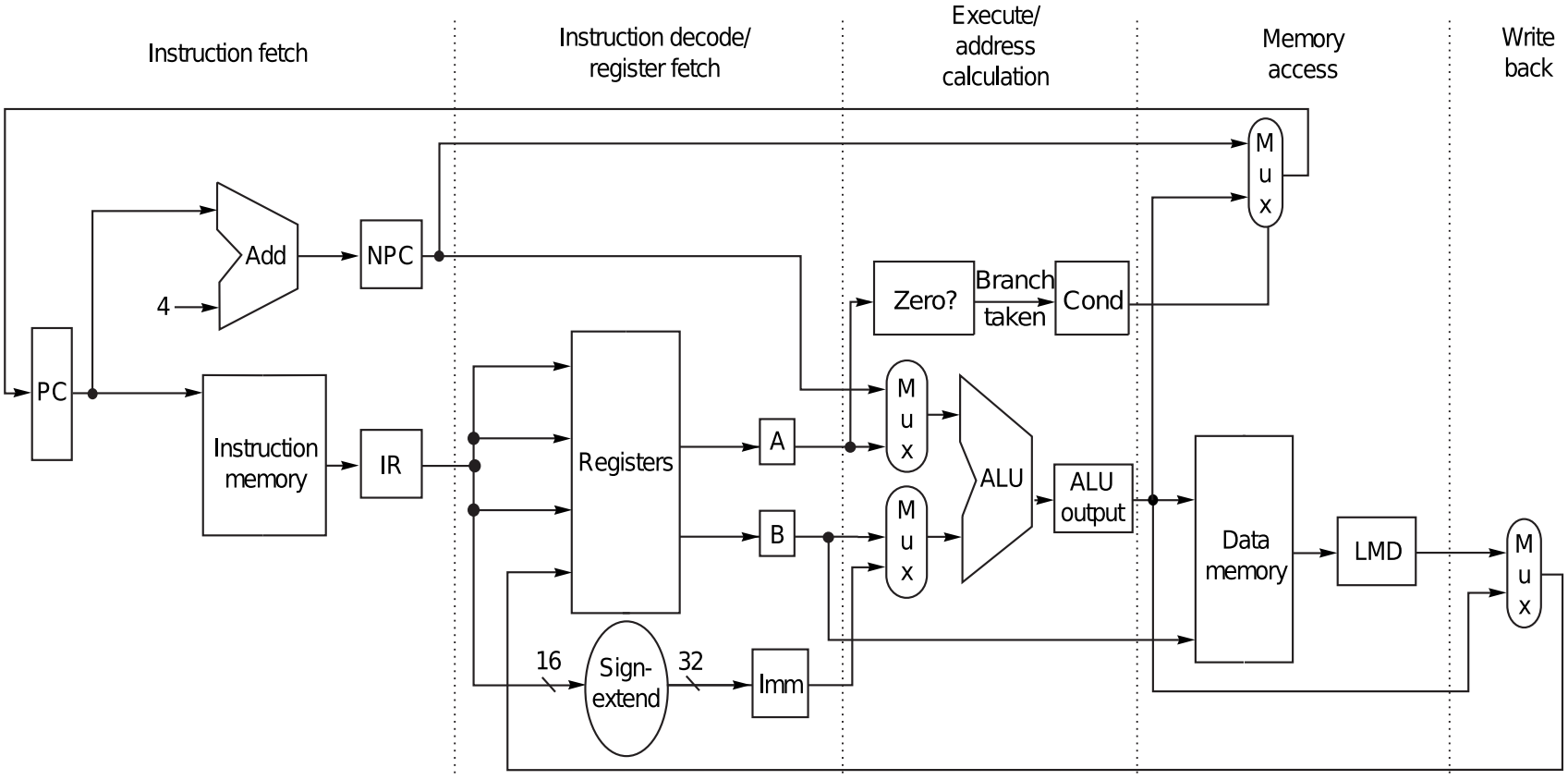


MAIN COMPONENTS OF THE TWIN-REGENERATOR GAS TURBINE:
(A) accessory drive; (B) compressor; (C) right regenerator rotor; (D) variable nozzle unit; (E) power turbine; (F) reduction gear; (G) left regenerator rotor; (H) gas generator turbine; (I) burner; (J) fuel nozzle; (K) igniter; (L) starter-generator; (M) regenerator drive shaft; (N) ignition unit.

The Disadvantage of Turbine Cars: Acceleration

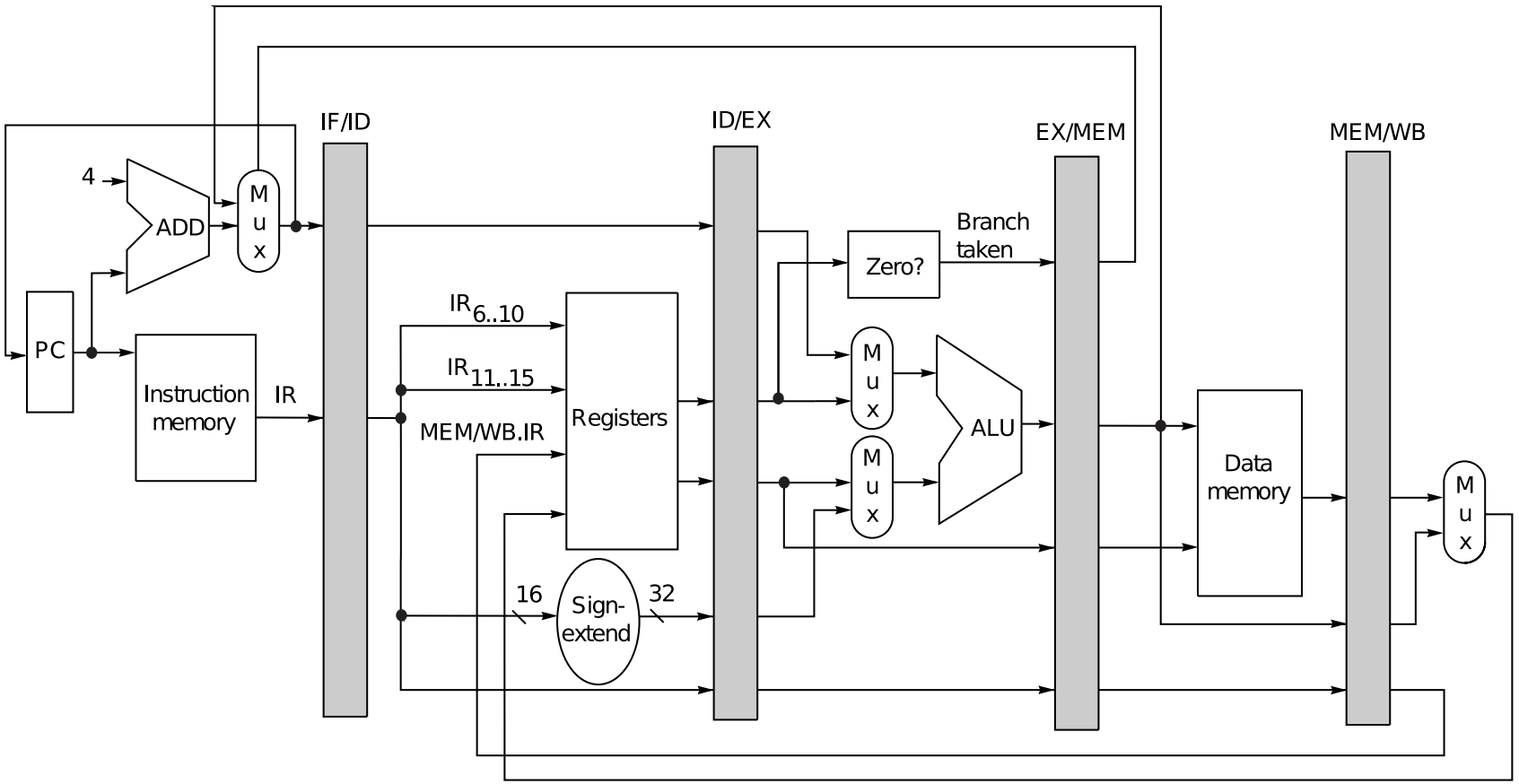
When can be just as important as *what*

Processor Design 101



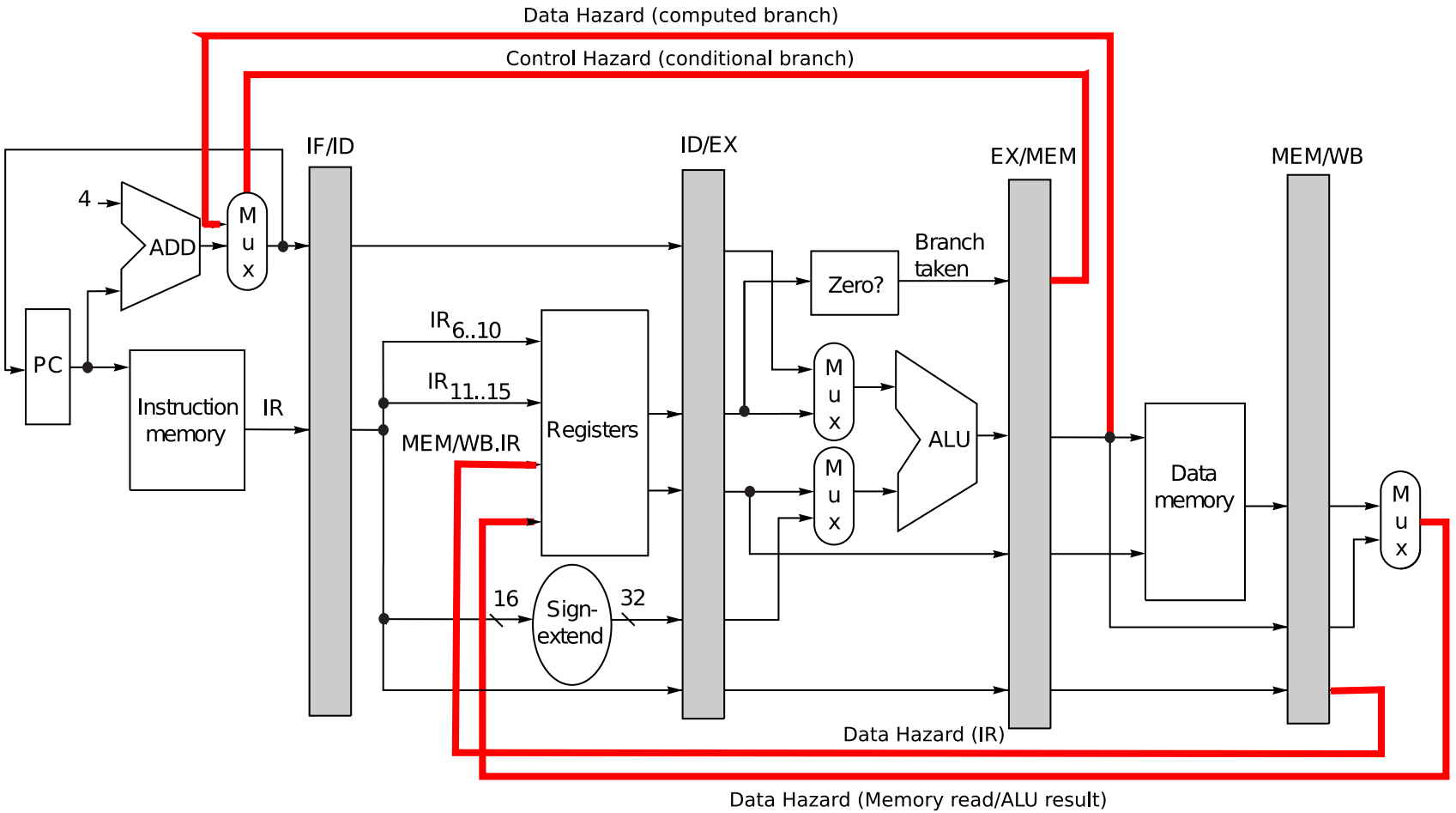
Hennessey and Patterson, *Computer Architecture: A Quantitative Approach*, 2007.

Pipeline It!



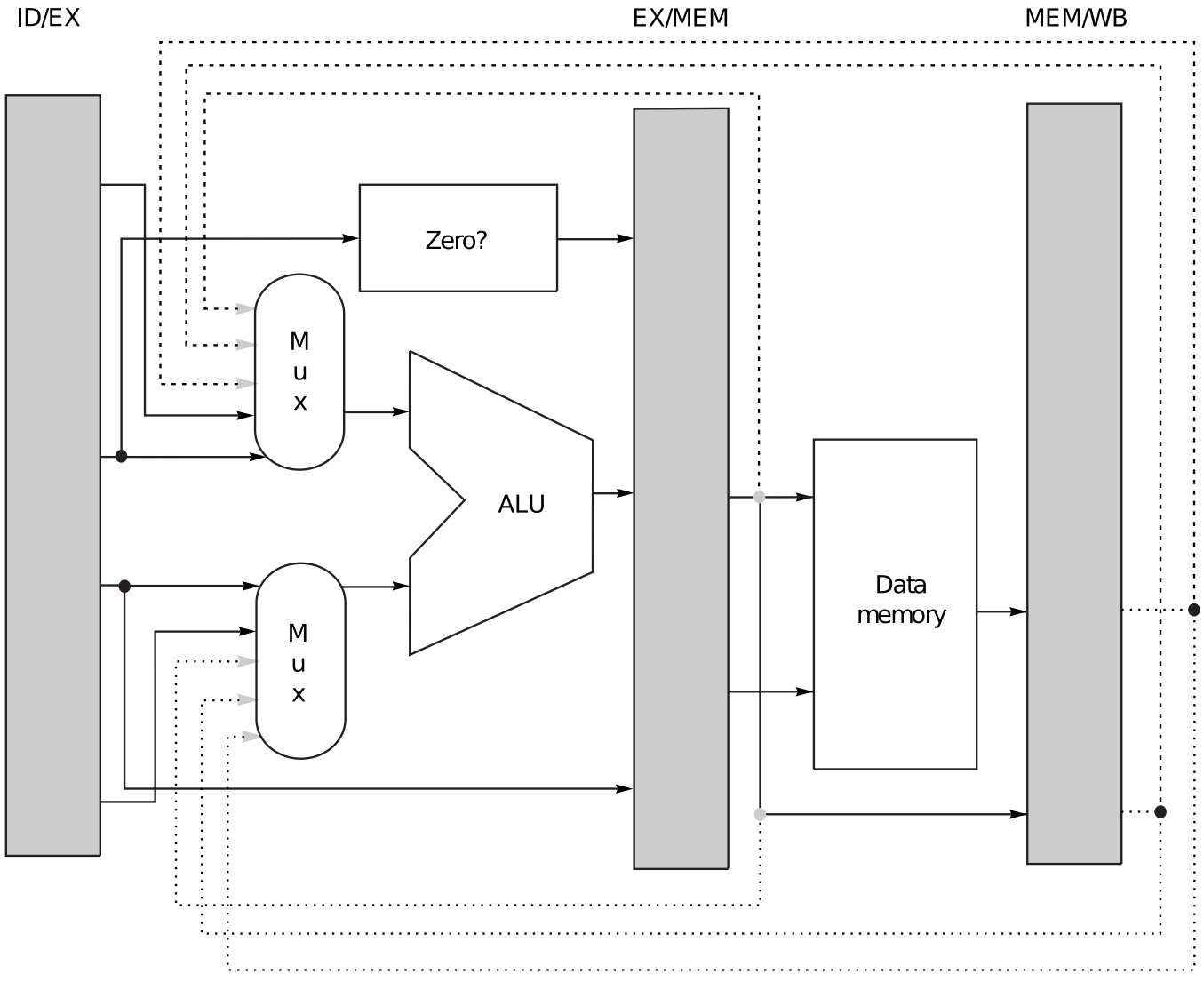
Hennessey and Patterson, *Computer Architecture: A Quantitative Approach*, 2007.

Great Except for Hazards



Hennessey and Patterson, *Computer Architecture: A Quantitative Approach*, 2007.

Forwarding Can Reduce the Need to Stall...



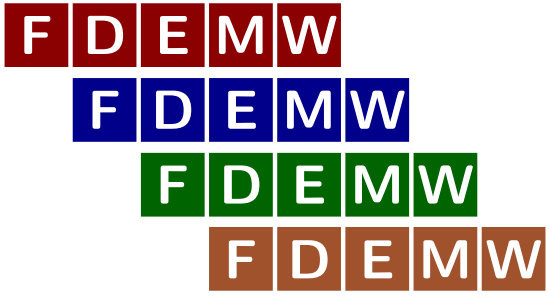
Hennessey and Patterson, *Computer Architecture: A Quantitative Approach*, 2007.

...But It Does Not Solve Everything...

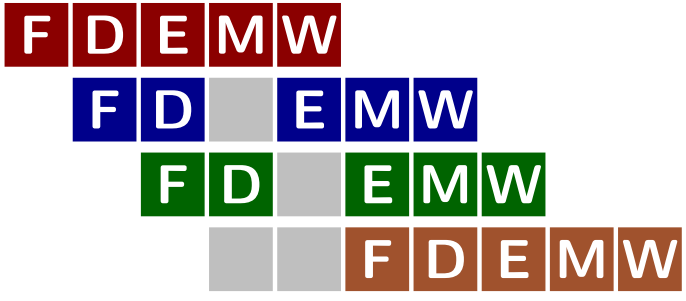
```
LD R1, 45(r2)
DADD R5, R1, R7
BE R5, R3, R0
ST R5, 48(R2)
```

Unpipelined

The Dream



The Reality

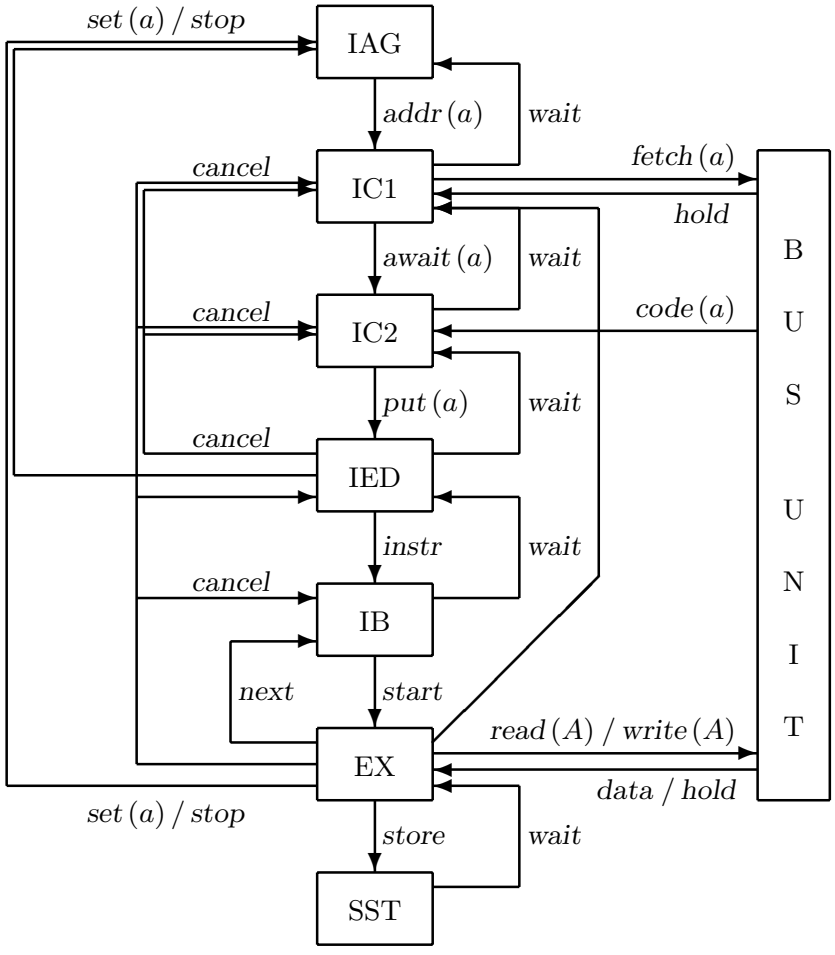


Memory Hazard

Data Hazard

Branch Hazard

...And It Makes Pipelines Complex



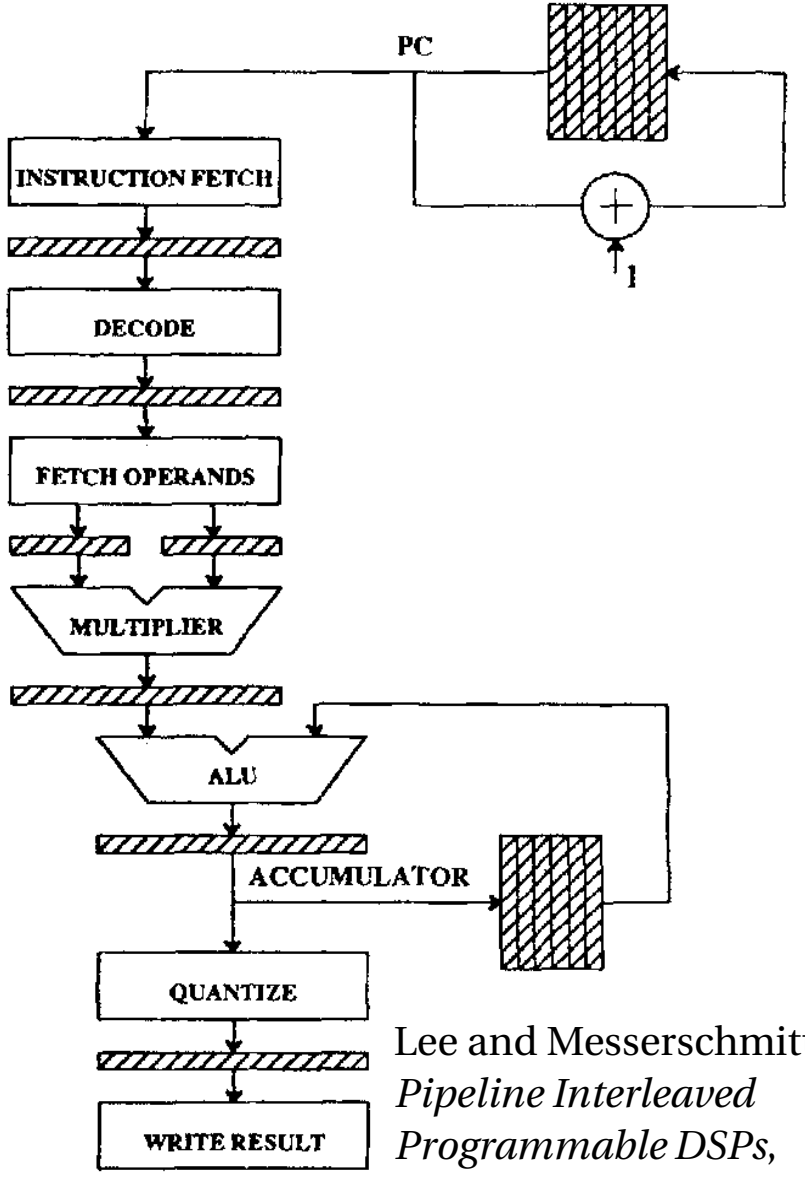
Motorola Coldfire pipeline from Ferdinand et al., Reliable and precise WCET determination for a real-life processor, EMSOFT 2001

Our Solution: Thread-Interleaved Pipelines

+

An old idea from the 1960s

But what about memory?



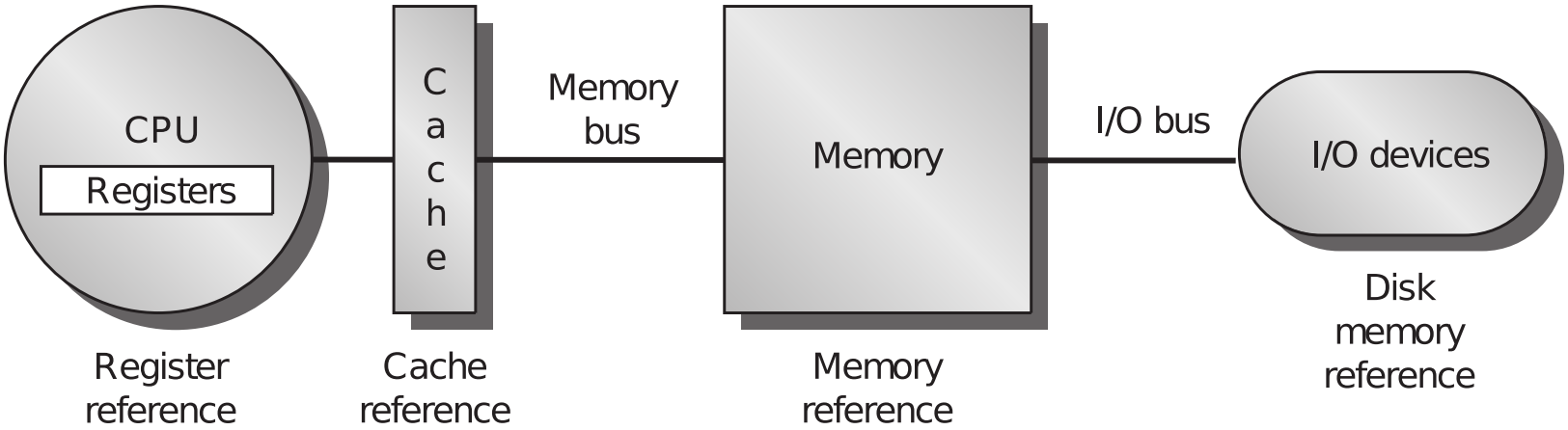
Lee and Messerschmitt, *Pipeline Interleaved Programmable DSPs*, ASSP-35(9), 1987.

Memory Hierarchy is an Old Idea

> *Ideally one would desire an indefinitely large memory capacity such that any particular ... word ... would be immediately available. We ... recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.*

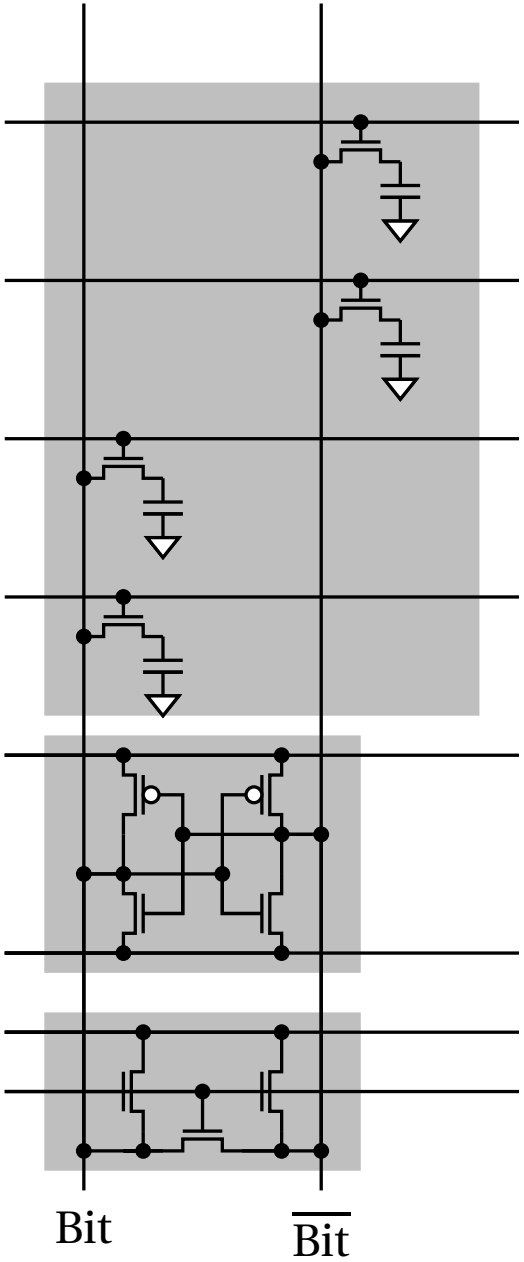
— Arthur W. Burks, Hermann H. Goldstine, and John von Neumann, *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument*, 1946

Memory System Design 101



Hennessey and Patterson, *Computer Architecture: A Quantitative Approach*, 2007.

DRAM Circuit Operation

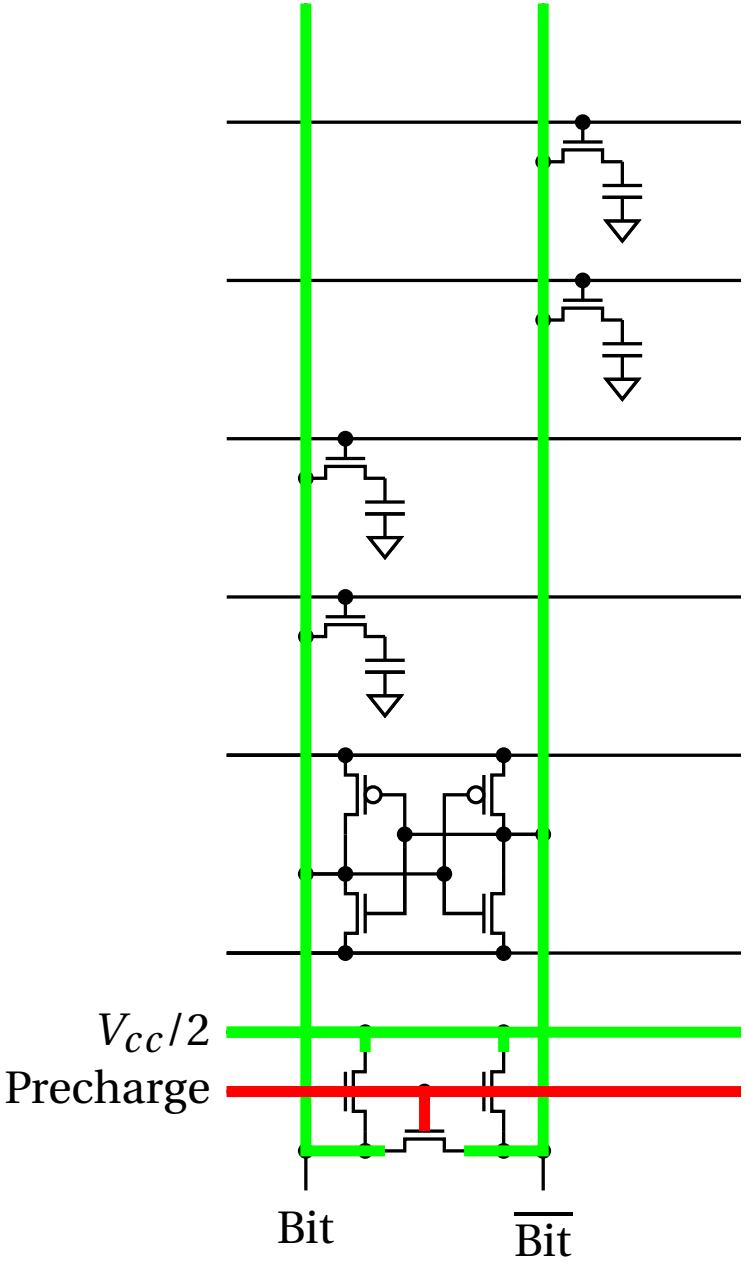


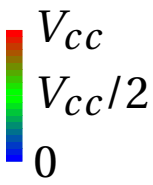
Bit Cells

Sense Amplifier

Equalization Circuit

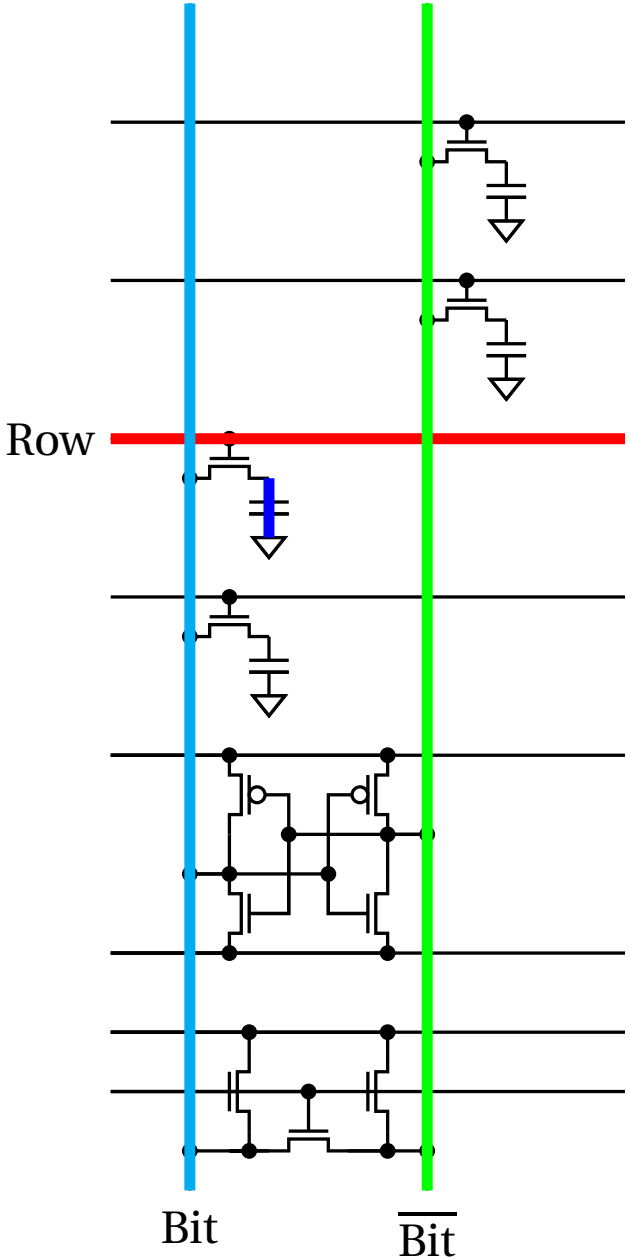
DRAM Circuit Operation

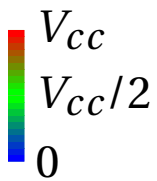




- **Precharge**
  Set bit lines to $V_{cc}/2$

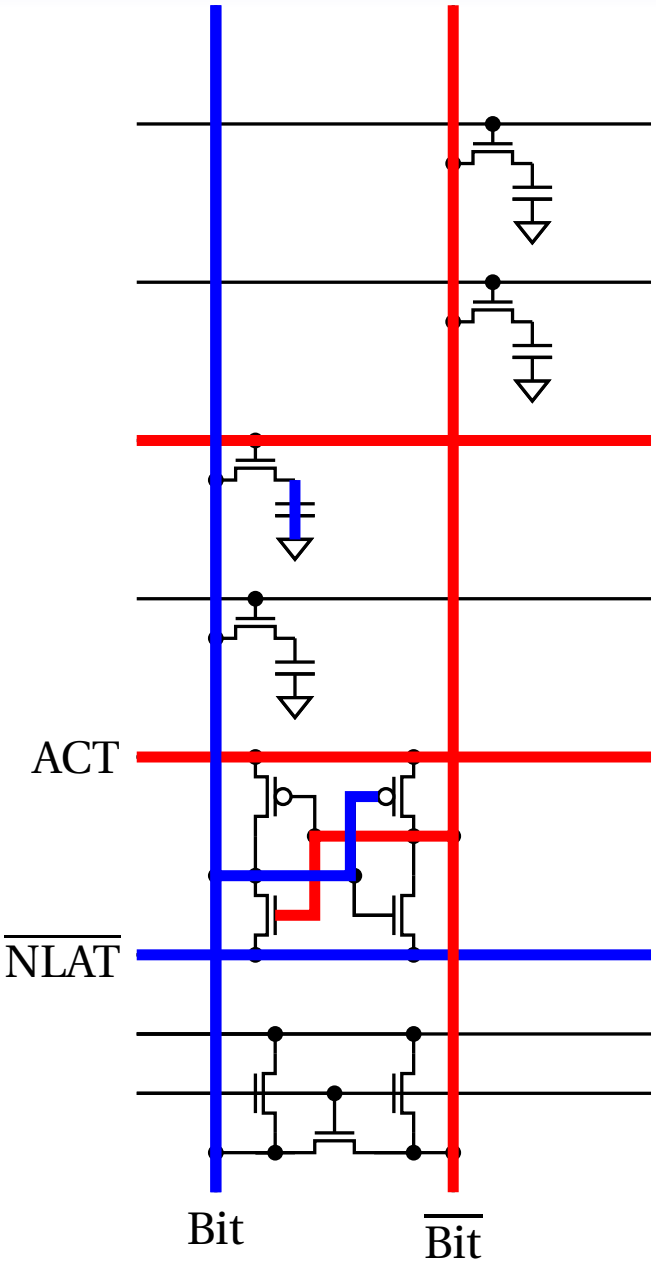
DRAM Circuit Operation





- **Precharge**
  Set bit lines to $V_{cc}/2$
- **Access**
  Select a row; perturb bit lines

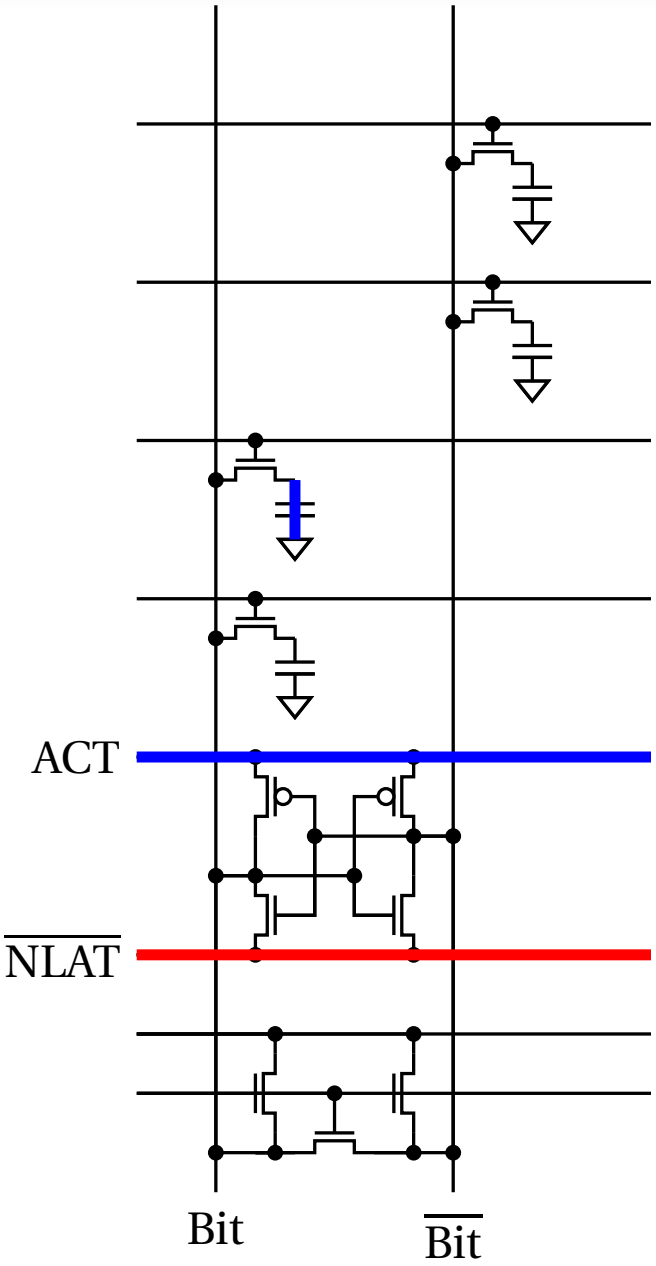
DRAM Circuit Operation



V_{cc}
$V_{cc}/2$
0

- **Precharge**
  Set bit lines to $V_{cc}/2$
- **Access**
  Select a row; perturb bit lines
- **Sense**
  Enable sense amplifier
  Drive bit lines to rails

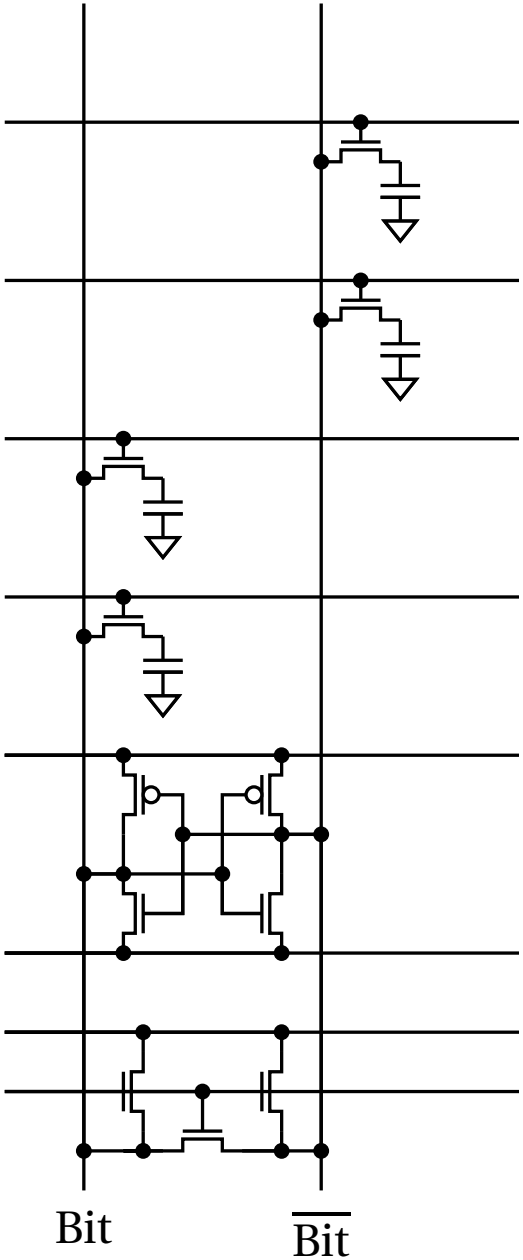
DRAM Circuit Operation

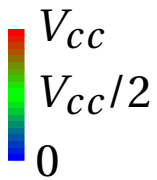


V_{cc}
$V_{cc}/2$
0

- **Precharge**
  Set bit lines to $V_{cc}/2$
- **Access**
  Select a row; perturb bit lines
- **Sense**
  Enable sense amplifier
  Drive bit lines to rails
- **Restore**
  Disable sense amplifier

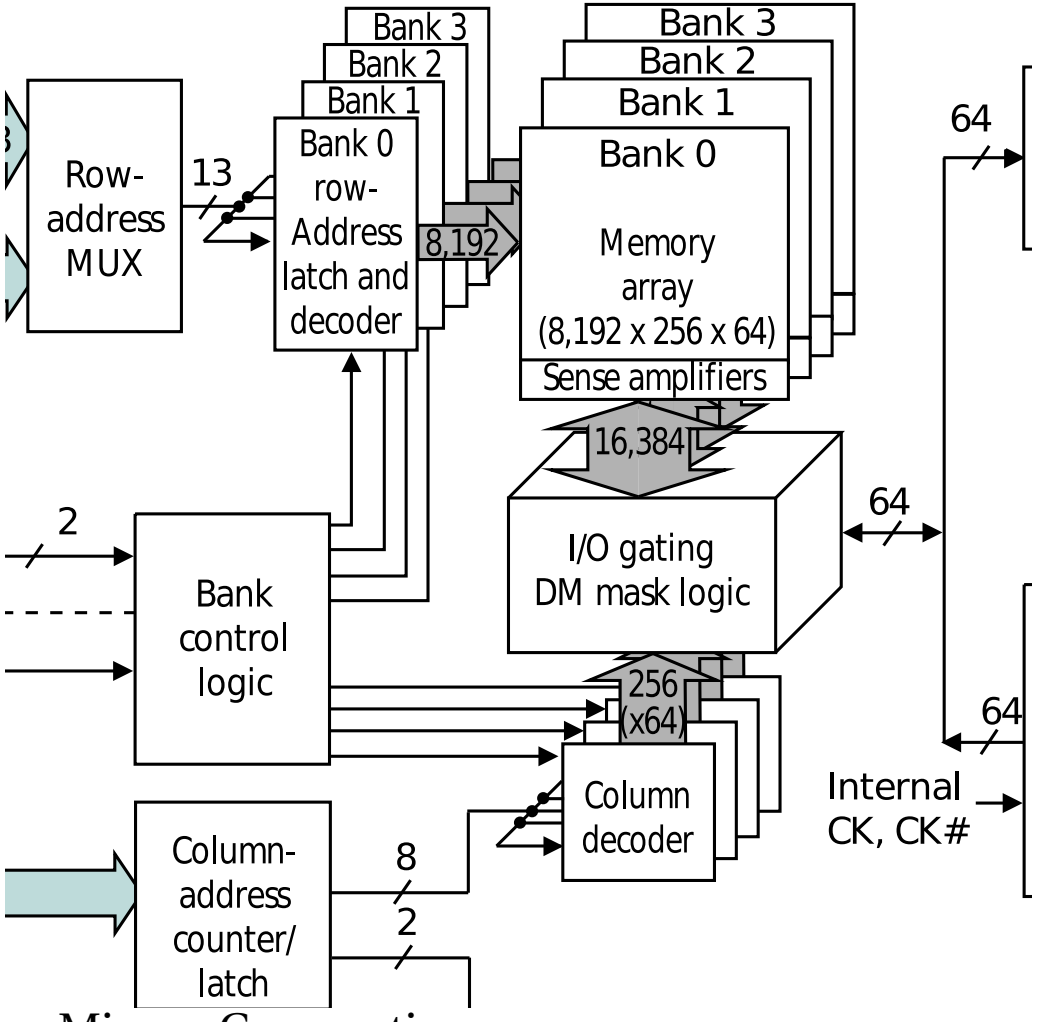
DRAM Circuit Operation

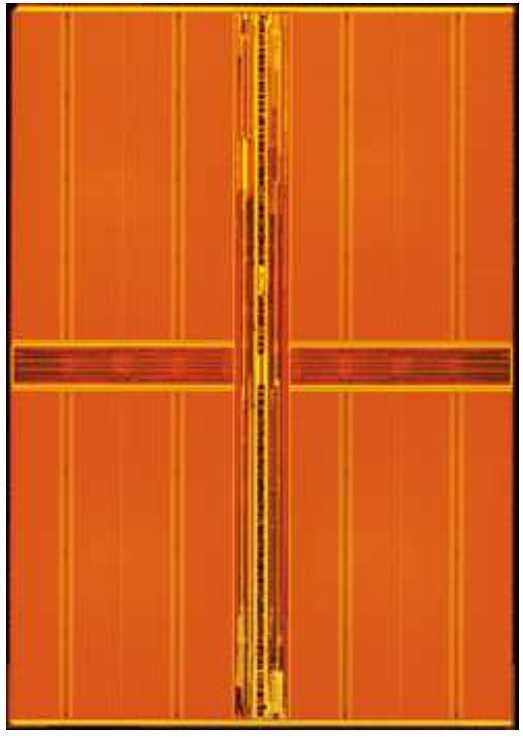




- **Precharge**
  Set bit lines to $V_{cc}/2$
- **Access**
  Select a row; perturb bit lines
- **Sense**
  Enable sense amplifier
  Drive bit lines to rails
- **Restore**
  Disable sense amplifier

Modern DRAMs Have Banks

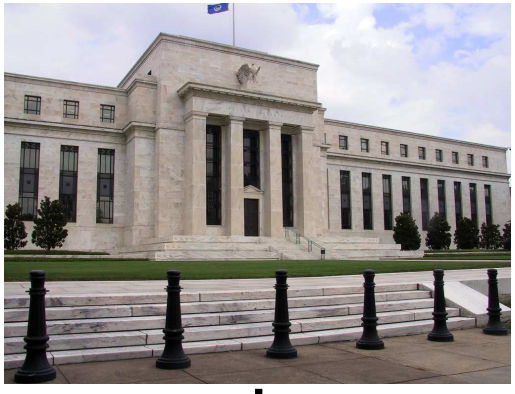


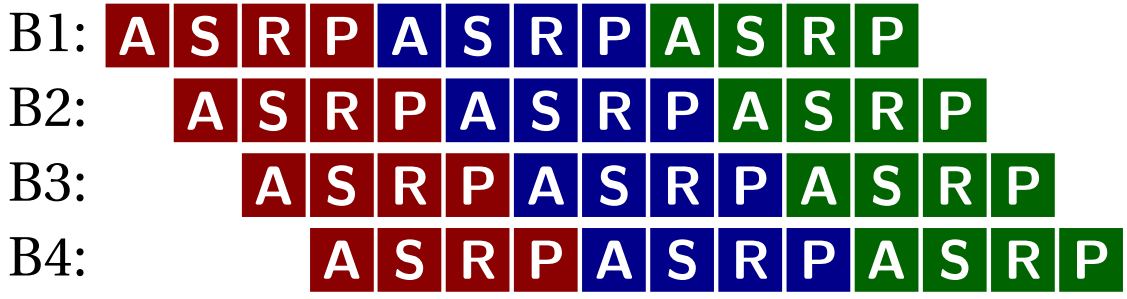


DDR2: 4–8 pipelined banks

DDR3: 8+ pipelined banks

Banks Enable Pipelining



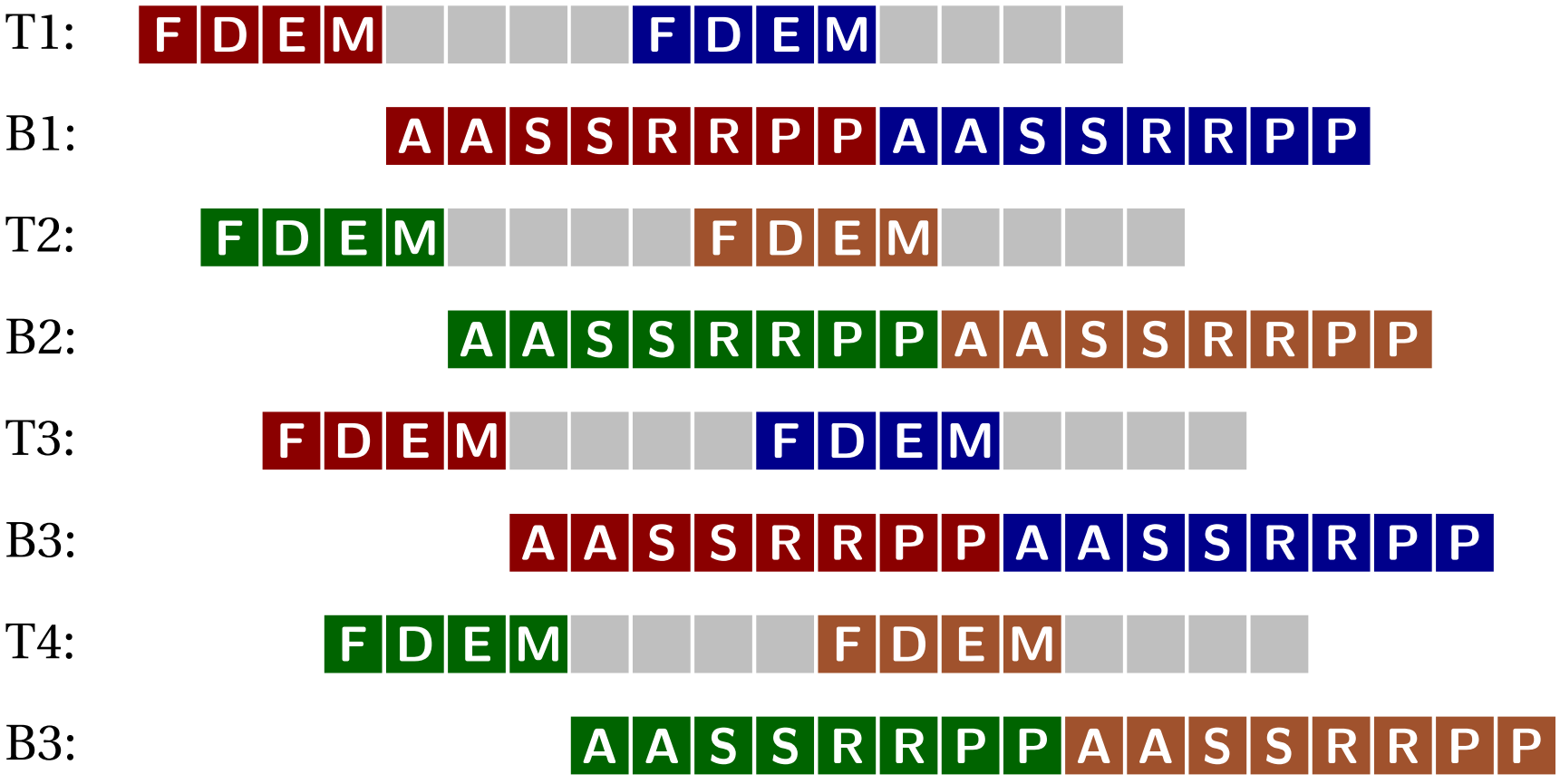


Not to scale!

Our Solution: Pipeline-Synchronized Memory

- One  per 
- Tune processor  frequency to memory

Pipeline-Synchronized Memory



Pipeline-Synchronized Refinements

-  Add scratchpad memories
-  Use round-robin scheduling for sharing banks
-  Add ranks and DIMMs for more parallelism
-  Get easier-to-pipeline memory

Conclusions

- PRET goal: predictable, temporally isolated threads
- Thread-interleaved pipelines avoid hazards
- Pipelined DRAMs allow elegant sharing of resources
- Pipeline-synchronized memory hierachy
- Working on an FPGA prototype